Looking for data: Information seeking behaviour of survey data users

Dissertation
zur Erlangung des akademischen Grades

Doctor philosophiae (Dr. phil.)

eingereicht

an der Philosophischen Fakultät
der Humboldt-Universität zu Berlin

von Tanja Friedrich

Die Präsidentin der Humboldt-Universität zu Berlin
Prof. Dr.-Ing. Dr. Sabine Kunst

Die Dekanin der Philosophischen Fakultät
Prof. Dr. Gabriele Metzler

Gutachterinnen

Erstgutachterin:	Prof. Vivien Petras, PhD
Zweitgutachterin:	Prof. Dr. Alexia Katsanidou
Drittgutachterin:	Prof. Dr. Elke Greifeneder

Datum der Disputation: 12. Juni 2020

Abstract

From information behaviour research we have a rich knowledge of how people are looking for, retrieving, and using information. We have scientific evidence for information behaviour patterns in a wide scope of contexts and situations, but we don't know enough about researchers' information needs and goals concerning the usage of research data. Having emerged from library user studies, information behaviour research especially provides insight into literature-related information behaviour. This thesis is based on the assumption that these insights cannot be easily transferred to data-related information behaviour. In order to explore this assumption, a study of secondary data users' information-seeking behaviour was conducted. The study was designed and evaluated in comparison to existing theories and models of information-seeking behaviour. The underlying research paradigm is social constructivism.

The overall goal of the study was to create evidence of actual information practices of users of one particular retrieval system for social science data in order to inform the development of research data infrastructures that facilitate data sharing, which is a vital demand of international information infrastructure policy. The empirical design of this study follows a mixed methods approach; more precisely, an exploratory sequential design was applied. This includes a qualitative study in the form of expert interviews and – building on the results found therein – a quantitative web survey of secondary survey data users.

The core result of this study is that community involvement plays a pivotal role in survey data seeking. The analyses show that survey data communities are an important determinant in survey data users' information seeking behaviour and that community involvement facilitates data seeking and has the capacity of reducing problems or barriers. Survey data communities emerge and persist, because knowledge about survey data is handed down from senior researchers to junior researchers or shared between peers. Community involvement increases with growing experience, seniority, and data literacy.

In line with social constructivist aims of inquiry, this study's contribution to research is twofold. In theoretical respect, it advances information behaviour research by modelling specific information seeking behaviour. The model of data users' information seeking

behaviour is presented in a diagram that is based on the primary findings of the study. In practical respect, the study specifies data-user oriented requirements for systems design.

Zusammenfassung

Die Informationsverhaltensforschung liefert zahlreiche Erkenntnisse darüber, wie Menschen Informationen suchen, abrufen und nutzen. Wir verfügen über Forschungsergebnisse zu Informationsverhaltensmustern in einem breiten Spektrum von Kontexten und Situationen. Jedoch wissen wir bis heute nicht genug über die Informationsbedürfnisse und Ziele von Forscherinnen und Forschern bei der Nutzung von Forschungsdaten. Die Informationsverhaltensforschung, die aus Bibliotheksnutzungsstudien hervorgegangen ist, gibt insbesondere Aufschluss über das literaturbezogene Informationsverhalten. Die hier vorgelegte Arbeit basiert auf der Annahme, dass sich diese Erkenntnisse nicht ohne weiteres auf das datenbezogene Informationsverhalten übertragen lassen. Um diese Annahme zu überprüfen, wurde eine Studie zum Informationsverhalten von Sekundärnutzerinnen und -nutzern von sozialwissenschaftlichen Forschungsdaten durchgeführt. Die Studie wurde vor dem Hintergrund bestehender Theorien und Modelle des Informationssuchverhaltens konzipiert und ausgewertet. Die Untersuchung orientiert sich am Forschungsparadigma des Sozialkonstruktivismus (*social constructivism*).

Das übergeordnete Ziel der Studie war es, Erkenntnisse zur tatsächlichen Informationspraxis der Nutzerinnen und Nutzer eines bestimmten Retrievalsystems für sozialwissenschaftliche Daten zu erlangen, um die Entwicklung von Forschungsdateninfrastrukturen zu unterstützen, die den Datenaustausch erleichtern sollen. Damit bedient diese Untersuchung eine wichtige Forderung der internationalen Informationsinfrastrukturpolitik. Das empirische Design dieser Studie folgt einem Mixed-Methods-Ansatz. Die Untersuchung folgt einem explorativ sequentiellen Design. Dieses beinhaltet eine qualitative Studie in Form von Experteninterviews und – darauf aufbauend – eine quantitative Studie aufgrund einer Online-Befragung von Sekundärnutzerinnen und -nutzern von Daten aus Bevölkerungs- und Meinungsumfragen (Umfragedaten).

Im Kern hat die Untersuchung ergeben, dass die Einbindung in die Forschungscommunity bei der Datensuche eine zentrale Rolle spielt. Die Analysen zeigen, dass Communities bei der Informationssuche der Nutzerinnen und Nutzer von Umfragedaten eine wichtige Determinante darstellen und dass die Einbindung in die Community die Datensuche erleichtert. Die Einbindung in die Community hat das Potential, Probleme oder Barrieren bei der Datensuche zu reduzieren. Umfragedaten-Communities entstehen und bestehen, weil

das Wissen über Umfragedaten von erfahrenen Nutzerinnen und Nutzer an Nachwuchsforscherinnen und -forscher weitergegeben oder innerhalb von Peergroups geteilt wird. Die Einbindung in die Community nimmt mit zunehmender Erfahrung, Seniorität und Datenkompetenz (*data literacy*) zu.

In Übereinstimmung mit sozialkonstruktivistischen Untersuchungszielen leistet diese Studie einen doppelten Beitrag zur Forschung. In theoretischer Hinsicht bringt sie die Forschung zum Informationsverhalten durch die Modellierung des Datensuchverhaltens voran. Das auf den primären Ergebnissen der Studie basierende Modell des Informationssuchverhaltens der Datennutzerinnen und -nutzer wird diagrammatisch dargestellt. In praktischer Hinsicht liefert die Studie Empfehlungen für das Design von Dateninfrastrukturen, basierend auf empirischen Anforderungsanalysen.

Acknowledgements

I would like to thank all the people who have supported and advised me during this research project. First and foremost, I am very grateful to the reviewers Professor Vivien Petras, Professor Alexia Katsanidou, and Professor Elke Greifeneder for their detailed and insightful reviews. I would also like to thank my team leader at GESIS, Reiner Mauer, who provided me with the freedom and the working environment that is necessary to carry out an extensive research project like this. A number of other researchers and colleagues have supported and advised me in various ways, in particular: Dr. Insa Bechert, Dr. Libby Bishop, Dr. Marcus Eisentraut, Dr. Dagmar Kern, Dr. Katharina Kinder-Kurlanda, Dr. Anja Perry, Dr. Jonas Recker, Dr. Pascal Siegers, Oliver Watteler, and Wolfgang Zenk-Möltgen. Many other colleagues from GESIS have shown me their support in many different ways, which I am very thankful for. Colleagues and friends have participated in pretests, expert reviews, and respondent debriefings for the qualitative and quantitative study. I thank every one of them for their generous participation and valuable feedback. Furthermore, this research could not have been done without the support of the participants who have kindly agreed to take part in the expert interviews and the many people who have completed the survey on data seeking. I am beyond grateful for all their contributions. I am also grateful that they all allowed me to publish their contributions through a data archive and share them with the research community.

Finally, I would like to thank my family and friends for their support and companionship. I thank my husband Markus Friedrich for always being there for me and for always believing that I could do this. I thank my son Konrad for his positive energy that keeps reminding me of the most important things in life. I am thankful to my parents, Ulrike and Hans Friedrich, for having supported and encouraged me all my life.

I dedicate this work to my daughter Jule, who has taught me the two most important things: how to carry on and how to let go.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AAPOR *American Association of Public Opinion Research*

CILS4EU *Children of Immigrants Longitudinal Survey in four European Countries*

DANS *Data Archiving and Networked Services*

DBK *Datenbestandskatalog*

DDI *Data Documentation Initiative*

DEM *Documentation Evaluation Model for Social Science Data*

DISISS *Design of Information Systems in the Social Sciences*

DOI *Digital Object Identifier*

EU *European Union*

EUDAT *European Data Infrastructure*

FAIR *Findable, Accessible, Interoperable, Re-usable*

GESIS *GESIS Leibniz Institute for the Social Sciences*

GMF *Gruppenbezogene Menschenfeindlichkeit*

ICPSR *Inter-university Consortium for Political and Social Research*

INFROSS *Information Requirements of the Social Sciences*

INISS *Information Needs and Information Services*

LIS *Library and Information Science*

PC *Personal Computer*

Pew ATP *Pew American Trends Panel*

PI *Principal Investigator*

RatSWD *Rat für Sozial- und Wirtschaftsdaten*

RR1 *Response Rate 1 (AAPOR)*

RR2 *Response Rate 2 (AAPOR)*

XML *Extensible Markup Language*

A. Introduction

1. Problem Statement

In an envisioned research culture of data sharing, how can we be sure that navigating through the data deluge (Hey and Trefethen 2003; Marcum and George 2010) will lead secondary researchers to the appropriate data? Almost 30 years ago, economist Martin David, expert in public statistics and survey data analysis, expressed his concerns about data seeking with regard to successful data sharing: "Can shared data be as easily decoded as a shared library book?"(David 1991, 93) He suggested approaching the task of providing helpful data services by taking the users' perspective: "Affirmative answers […] require that we identify the expectations and needs of the secondary data user and that we provide support to meet those needs."(David 1991, 93)

Today it seems we haven't got far in that respect. Admittedly, it can be stated that, in social and economic sciences, a culture of data sharing has been achieved (Huschka et al. 2011). This is chiefly true for those researchers of political science and sociology who work mainly quantitatively. In form of central data archives they had adequate data sharing infrastructure ready since the 1960s (Jacoby 2010). However, we are still missing empirical evidence to identify the "expectations and needs of the secondary data user" mentioned by David. More recently it has been pointed out that "sharing of social science data […] has received inadequate attention from the information science community."(J. Niu and Hedstrom 2008, 1) Given that extensive research data infrastructures have not only been a vital demand in international research policy for years, (High level Expert Group on Scientific Data 2010; Kommission Zukunft der Informationsinfrastruktur 2011; Wissenschaftsrat 2011) but are already being developed (e.g., the EU-funded project EUDAT[1]), the need for research in this area seems even more urgent. In particular, the user perspective must not be neglected when large investments are to pay off (Zimmerman 2007). This implies that knowledge about secondary use of research data is crucial for the development of data infrastructures that are suited to make data sharing feasible. In the first and to this day most comprehensive user study of social science data archives, the author Kathleen Heim also named this reason as the main driver of her research: "Without an understanding of the role

[1] http://www.eudat.eu/, accessed October 5, 2020.

played by statistical and machine-readable data in the working life of the social scientist, any information system designed will fail to anticipate the total information needs of these disciplines." (Heim 1980, 21) An application-oriented Library and Information Science (LIS) is one of the disciplines that have to provide basic research to inform this development.

Scientific research has always been data-dependent. Only in the last few decades, through the definite advent of ubiquitous computerization, virtually all research has become computer-driven and, as a consequence, is producing and processing ever-more data (High level Expert Group on Scientific Data 2010). How to cope with the data deluge has become a vitally important question for researchers and funding agencies, resulting in raising awareness for the paradigm shift in research (Hey, Tansley, and Tolle 2009; Lynch 2009) and in calls for data sharing (Pilat and Fukasaku 2007). To the same extent to which e-science (or e-research) is influencing research culture and practice, it is also altering the business of research service providers, such as data centres and libraries. While on the one hand, major funding agencies encourage (if not commit) researchers to make their data available for replication and re-use (Deutsche Forschungsgemeinschaft 2015; Economic and Social Research Council 2018; National Science Foundation 2012), libraries and data centres on the other hand are expected to provide respective services for data stewardship, sharing and access (Wissenschaftsrat 2011). "Research data sets need to be discoverable and accessible in similar ways as publications are" (van der Graaf, Waaijers, and Davidson 2011, 7), is a request by the international network of funding agencies, Knowledge Exchange[2]. Research libraries in particular possess rich expertise in providing access to research information, and in the last decades, they have gradually met the challenge to help research infrastructure evolving into e-infrastructure. However, their expertise in mainly providing access to written information is being probed particularly with regard to the development of infrastructures for data sharing (Borgman 2010; Gold 2007; Palmer et al. 2009). Research datasets form a distinct, independent type of resource (van der Graaf, Waaijers, and Davidson 2011), which is assuming manifold shapes according to disciplines, research methods and usage scenarios (Blue Ribbon Task Force on Sustainable Digital Preservation and Access 2010; Borgman

[2] http://www.knowledge-exchange.info/, accessed October 5, 2020.

2012). Fundamental research on the usage of research data is needed to inform the development of information infrastructure for data-intensive scientific discovery.

While research libraries acknowledge their responsibilities with regard to data stewardship (Gold 2007; Borgman 2010), researchers in library and information science (LIS) increasingly investigate various aspects of data sharing (Jacoby 2010). But as far as the knowledge about "expectations and needs of the secondary data user" (David 1991, 93) is concerned, LIS is facing a research gap. One reason for the lack of research in this important area may be that, in the past, the curation of research data has largely taken place outside libraries, in specialized information centres (for the case of survey data: in social science data archives) and therefore has never been in the focus of LIS researchers. Nicholas Weber states that, even though the field of information science "has traditionally studied some of the most difficult problems in the use of large-scale information resources, including the meaningful organization, access, management and storage of scholarly products in all of their formats and encodings […] this space is already crowded with sociologists, economists, computer scientists and statisticians, to name a few of the disciplines involved." (Weber 2013, 23) To catch up on the shortcomings regarding research data services, Weber suggests that "we must better apply what we've traditionally known about citation behaviour, document retrieval and information seeking to a data-intensive paradigm, while simultaneously avoiding generic simplifications such as 'publications are just like datasets'." (Weber 2013, 23) Fortunately, with information behaviour research LIS has a relevant sub-discipline that offers a wide range of approaches to study data seeking. Information needs and uses as well as information seeking behaviour are research topics which have been thoroughly studied in the field. However, investigations of researchers' information behaviour have largely dealt with their seeking and using of literature. For this thesis, findings from these studies are expected to being only partly transferrable to the usage of datasets, but general theories, models and instruments of information behaviour studies are assumed to be applicable. Additional to LIS researchers' theoretical and empirical findings on information behaviour, this thesis will draw on data archives' knowledge from practice in data sharing to gain insight in secondary data users' information seeking behaviour. Combining both realms of knowledge is assumed to be a fruitful approach.

2. Area of Research

Present-day information behaviour research presents a rich knowledge on how different groups of people (e.g. students, professionals) interact with different types of information in different context and for different purposes (occupational, every day information use etc.). Having emerged from library user studies, the field dates back many decades (Case and Given 2016).

While early studies were focused on the use of specific information systems (usually libraries) with regard to an evaluation of these systems (Paisley 1965), later research looked more closely at characteristics of the systems' users (Case and Given 2016). This change of perspective, often referred to as *the user-centred turn*, is commonly associated with the 1986 paper of Brenda Dervin and Michael Nilan (Talja and Hartel 2007), but can already be surmised in user studies from the 1960s (Bates 2004). Over time, there has been a large increase of "information needs and uses" studies which in turn led to theoretical as well as methodological growth and – at least in the Anglosphere – eventually resulted in the establishment of information behaviour research as an important sub-discipline within information studies (Pettigrew et al. 2001). These developments have been studied and described prominently by Donald Case and Lisa Given (2016). For a comprehensive collection of theories and models of information behaviour in general a key source still is the book by Karen Fisher et al. (2005) even though there have been further developments and new trends, especially research approaches employing evolutionary and developmental theoretical frameworks. More recently, Amanda Spink and Jannica Heinström have collected papers on "cutting-edge" developments in information behaviour research that are based on "evolutionary and developmental foundations, meta-synthesis, individual and contextual dimensions, information interaction, impact of information and longitudinal process models" (Spink and Heinström 2011, XVII).

The mentioned books give a good overview of the large amount of research that has been done in the field, but of course they do not investigate every one of the thousands of studies that have been conducted. In 2012, Donald Case deemed it "easily possible" that there were more than 10,000 publications on "information needs, uses, seeking, and other aspects of information behavior" (Case 2012, 277). From this abundance, only the most important works which have a closer connection to the topic investigated here will be presented in the

following. Mainly these are studies of information behaviour of researchers and, more precisely, of social scientists.

Since the study of students' and researchers' information behaviour has been in the focus of the discipline from the beginning (Leckie 2005) and still dominates the field in today (Borgman 2007), we have plenty of knowledge and empirical evidence in this area. In the beginning, the research has been largely limited to scientists and engineers (Case and Given 2016; Wilson 2000). The social sciences were the next discipline to be studied (Gannon-Leary, Bent, and Webb 2007), the first investigations probably being the Project on Scientific Information Exchange in Psychology, undertaken by the American Psychological Association, beginning in 1961 (Paisley 1965). The first major study of social scientists' information behaviour was conducted in the UK from 1967 to 1971: the INFROSS[3] study, followed by DISISS[4] and INISS[5] (1975-1980). The 1980s saw a decrease in investigations of social scientists' information behaviour, maybe because the big studies in the 1970s had a saturating effect (Slater 1988). Another reason for the decline could be that due to rapid developments in information technology, information professionals believed that problems in information provision could soon be solved by the sole employment of proper technology (Janes 2009). By the end of the 1980's, a seminal empirical study aimed at generating concrete guidance for information retrieval system design from behavioural characteristics of the information-seeker. This was the often cited study by David Ellis (1989), which led to a frequently adopted model of information seeking characteristics of social scientists. This model has since been modified and extended by Ellis himself (with regard to other research fields) (Ellis, Cox, and Hall 1993) and by other authors (e.g. Meho and Tibbo 2003).

It is apparent that, in the course of time, studies of "information needs and uses" of social scientists have evolved into "information behaviour studies" along with the general developments in information behaviour research. The behavioural study of Ellis can be seen as a landmark in this progress. Even though afterwards no major studies of information behaviour in the social sciences have been conducted, researchers of the smaller

[3] INFROSS = Information Requirements of the Social Sciences; for an overview of the project cf. Line (1971).
[4] DISISS = Design of Information Systems in the Social Sciences.
[5] INISS = Information Needs and Information Services; a study of communication and information flows in local authority social services departments.

investigations have since adopted the user-centred perspective and have increasingly based their studies on theories and models of information behaviour.

Unfortunately, we cannot be sure that the acquired knowledge from the numerous studies of social scientists' information behaviour is also applicable to their seeking for survey data. This is because, even though the need to access and use datasets has been reported, the investigation of information seeking behaviour of social scientists has always been focused on literature. Kathleen Heim, who conducted the to this day most comprehensive user study of social science data archives, stated in 1980 that, there had been "numerous studies of the information seeking behavior of social scientists" (Heim 1980, 1) aiming at improving library services; but even though part of these studies found evidence for the need of research data as an information source, they didn't lead to thorough investigations of needs and uses concerning social science data (Heim 1980). According to Heim, the primary reason for excluding the usage of data in further library studies was the fact that data stewardship was none of their services, but was carried out by data archives. She concludes that as a result, even major user studies such as INFROSS "failed to inquire about the use of the types of information found in data archives" (Heim 1980, 22).

The missing investigation of data user's information needs and behaviour correlates with the fact that these resources have been – and still are – less accessible than literature (Gould and Handler 1989). Further investigations in the 1980s arrived at similar conclusions: John Fletcher stated in 1982 that the demand for as well as the supply of statistical data in the field of economics were rising, but appropriate information systems were lacking (Flechter 1982). For the field of social policy and administration Colin Harris found in the same year that "data generated from earlier studies" (Harris 1982, 44) were only partly available and (thus) under-used. Another relevant empirical study, based on interviews and consultations with 73 individuals from economics, political science, sociology, psychology, and anthropology, which was published in 1989, still concluded: "Computer files of all types – from large data bases to smaller data files – are the staff of life for increasing numbers of social scientists. It is ironic that, […] these important sources of research information are in many cases difficult or impossible to obtain."(Gould and Handler 1989, 52)

Heim's 1980 study on users of data archives was, as already pointed out, a first and broad attempt to bring together the realms of library user studies and data archives. Corresponding with the state of the art in user studies at the time, she investigated the amount of users, their disciplines and their motivations (Heim 1980). Her findings are valuable descriptions of users, but not of their data seeking behaviour, as it is the aim of this thesis. Further relevant LIS research on usage of social science data has been conducted in the following two decades. In 1997, Carol Hert and Gary Marchionini published their findings on "Seeking Statistical Information in Federal Websites", an extensive usability evaluation of three websites, implementing multiple empirical methods. They investigated the types of users, their tasks or "statistical needs", and their strategies for finding information (Hert and Marchionini 1997). In the 2000s Jinfang Niu and Margaret Hedstrom developed and tested a "Documentation Evaluation Model for Social Science Data" with the goal of overcoming inadequate documentation of research data (J. Niu 2009; J. Niu and Hedstrom 2008, 2009). By asking social science researchers to judge documentation of data, they identified relations between user characteristics, the nature of data, and perceived documentation quality in terms of sufficiency and ease-of-use.

Research about data seeking behaviour is limited to very few studies that are concerned with particular problems, mostly related to information retrieval and usability issues. Not only do we need more research; we also need to take into account current practices of data-driven research as well as theoretical and empirical findings from studies in information seeking behaviour. The present study is aimed at the development of a model of data seeking behaviour in order to shed light on this previously understudied and undefined concept. In the present study, the concept of data seeking behaviour refers to information seeking behaviours that are directed towards a specific type of information resource. As a working definition, data seeking behaviour is understood as behaviours and practices that occur if people are looking for data that they can use to accomplish their work tasks.

3. Purpose Statement and Research Question

The overall goal of the study was to create empirical evidence for information-seeking behaviour patterns of social science data researchers in order to inform the development of

research data infrastructures that facilitate data sharing. The specific research question that this study investigates is:

What are the characteristics of researchers' information seeking behaviour with regard to survey data?

Due to her social constructivist perspective, the author is especially interested in how these characteristics and practices depend on social, interactive and contextual parameters. Therefore, the research question extends to influencing factors of survey data seeking. Guiding questions on characteristics and practices of people who are looking for survey data and questions on influencing factors are:

- What patterns occur in data seeking practices/ behaviours?
- What individual characteristics do survey data users have?
- What are the (social, situational) contexts of survey data users?
- What needs do survey data users have, what goals do they try to reach and what purposes do they pursue?
- What are requirements of survey data users who want to find data for reuse?
- What problems do survey data users encounter when looking for data? How do they solve them?

These guiding questions provide the starting point for the literature review and theoretical assumptions that are given in chapter "B. Theoretical Perspective".

Since there is a distinct need for research on data-seeking practices in general, and since multidisciplinary infrastructure solutions are desirable, it would seem to be appropriate to not restrict the investigation to social scientists. But, as already stated, unlike written documents, research data vary broadly according to disciplines, research methods and usage scenarios. Therefore, “the ways of and conditions for access to research data must be developed separately for the individual scientific disciplines […]” (Alliance of German Science Organisations 2010).This discipline-specific approach fits the social constructivist perspective in that it considers the individuals within “the world in which they live and work” (Creswell 2013, 24). For practical reasons, further restrictions to the research subject have to be made: The users will not be “social scientists” as a whole, since this is a population that is hardly to

be studied representatively. Instead, the population in focus will be users of social science data archives. Consequently, the specific type of research data will be the one that is available in these institutions: quantitative survey data. It seems to be fruitful to approach data users' information behaviour from the perspective of the users of social science data archives, since these are institutions, where data services are long established (cf. Nielsen and Hjørland 2014). From this starting point, we can benefit from the social science data archives' rich experience in serving secondary data users and probably make use of this knowledge for data services in other disciplines.

The research contribution of this thesis is twofold: Firstly, based on established knowledge, theoretical assumptions and empirical findings, the study resulted in a model of information-seeking behaviour with regard to research data and thus contributes to information behaviour theory. Secondly, the study aims at informing library and documentation practice by deducing concrete recommendations for infrastructure development.

4. Methodology

Research in LIS uses a broad variety of methods (Powell and Connaway 2004). Even though there are a few genuine LIS methods, above all bibliometric analyses, the discipline heavily draws on methods that arose from other fields. Which methods prevail in a specific LIS field depends on the adopted research paradigms. The social science perspective is very common in LIS research in general and in information behaviour research in particular (Ellis 2011). Since the social sciences are chiefly interested in observing human behaviour, empirical methods prevail in social research fields. LIS research has largely been influenced by these approaches and empirical social scientific methods are applied frequently and fruitfully in the field (cf. Dahinden 2013). Information behaviour researchers in particular favour survey methods (postal, web, and e-mail surveys), as has been shown in literature analyses (Case and Given 2016). Further empirical methods used by information behaviour researchers include: case studies; laboratory experiments; field experiments; brief interviews; intensive interviews; focus group interviews; network analyses; discourse analyses; diaries and experience sampling (cf. Case and Given 2016).

Despite the still prevailing popularity of survey research in LIS, it is apparent that qualitative methods have gained in importance since the late 1970s (Ellis 2011; González-Teruel and Abad-García 2012). This tendency is in accordance with the growth of qualitative research in the social sciences in general (Ellis 2011). The differentiation between quantitative and qualitative research paradigms is often made in the social sciences (Bryman 2012). It is a useful categorization of two different research approaches or strategies that have been seen as contradicting as well as complementing each other. Basically, quantitative research aims at testing theories deductively by scientific measurement of social reality (Bryman 2012). Qualitative research comprises inductive approaches that aim at theory generation by emphasising "the ways in which individuals interpret their social world" (Bryman 2012, 36). Given the need for conceptualization and theory generation in the relatively young field of LIS, researchers have repeatedly stressed the importance of applying qualitative methods (Ellis 2011; González-Teruel and Abad-García 2012). In particular, these approaches are deemed essential to understand key concepts such as *information need* (Ellis 2011).

However, Tom Wilson who was one of the early proponents of qualitative methods in information behaviour research later pointed out that the impending restriction to these approaches needed balancing by an interest in testing findings quantitatively (Wilson 2006). Wilson advocates "for multiple methods of research, rather than fixation with a single category of methods." (Wilson 2006, 681) David Ellis calls this approach of testing "a concern with empirical validation and exemplification" (Ellis 2011, 17) that he finds to be an element of "conceptual modelling in contemporary information behaviour research"(Ellis 2011, 17). This idea of combining qualitative and quantitative methods in a study is commonly known as *Mixed methods research* and has caught attention as a "third paradigm" in empirical research (Ma 2012, 1859). Ellis' diagnosis notwithstanding, combinations of quantitative and qualitative methods are not very common in LIS research (Fidel 2008). However, mixed methods approaches are increasingly advocated for, because they can "provide us with a richer understanding of information and information-related phenomena" (Ma 2012, 1866). This should be especially true for understudied phenomena such as data seeking.

The present study employs a mixed methods design for two reasons. First, a qualitative inquiry was carried out to reveal aspects of information behaviour and practices that could not be drawn from theoretical reasoning as well as to affirm those aspects that could be

identified theoretically. Second, a quantitative inquiry exemplifies the results from the qualitative study. There is a need to combine these two approaches, since using only one of them would be inadequate to help the understanding of the information seeking behaviour of survey data users (cf. Creswell 2014). By applying a mixed methods approach, this study both generates and tests theory on the subject (cf. Creswell 2014).

In particular, an *exploratory sequential design* is applied here (Creswell and Plano Clark 2011). This design puts the qualitative part before the quantitative part with the intention of developing a theory inductively, followed by a quantitative exemplification from this theory (Creswell 2014). The qualitative findings informed the formulation of hypotheses that were put to the test in the quantitative inquiry. John Creswell and Vicky Plano Clark (2011) specify this type of exploratory design as „the theory-development variant“ as opposed to the “instrument-development variant” (Creswell and Plano Clark 2011, 90). While the latter is prioritizing the quantitative phase of the study, the theory-development variant gives more importance to the qualitative findings by exemplifying them (Creswell and Plano Clark 2011).[6] This approach is more suited to the social constructivist viewpoint that is advocated here. Accordingly, the first investigation was aimed at “forming groups of attributes/themes” (Teddlie and Tashakkori 2009, 275) by application of a qualitative approach, while the second investigation is quantitative in that it provides “confirmatory statistical analysis” (Teddlie and Tashakkori 2009, 275). Following the description given by Charles Teddlie and Abbas Tashakkori (Teddlie and Tashakkori 2009), the empirical study proceeded in three steps: (1) A qualitative inquiry was used to identify constructs in the form of categories; (2) these categories were included in a questionnaire that was be presented to another population sample; (3) the resulting quantitative data was subject to construct validation by statistical analysis. The design of these three steps is further outlined in the following section.

[6] Similarly, Charles Teddlie and Abbas Tashakkori define mixed method design that involves „the process of *construct identification* and *validation*“ as *Typology development study* (Teddlie/Tashakkori 2009, p. 275, emphases in the original).

5. Research Design

The research design follows the tradition of conceptual modelling in information behaviour research, which is composed of "(1) the adoption of a social science perspective, (2) a qualitative as opposed to a quantitative orientation, (3) a focus on the modelling of information behaviour and (4) a concern with empirical validation and exemplification" as outlined by David Ellis (Ellis 2011, 17).

Accordingly, the research question is addressed with a mixed-methods design, combining qualitative interviews with a quantitative web survey. The leading part of the study is the qualitative study that was intended to primarily investigate the influencing factors of survey data seeking (independent variables). The characteristics and practices of survey data seeking (dependent variables) were studied in depth in the quantitative part. This approach allows for a comprehensive view on information seeking behaviour with regard to research data and leads to a model of data seeking that comprises macro and micro level information.

In concrete terms, the research proceeds as depicted in Figure 1 and explained below.

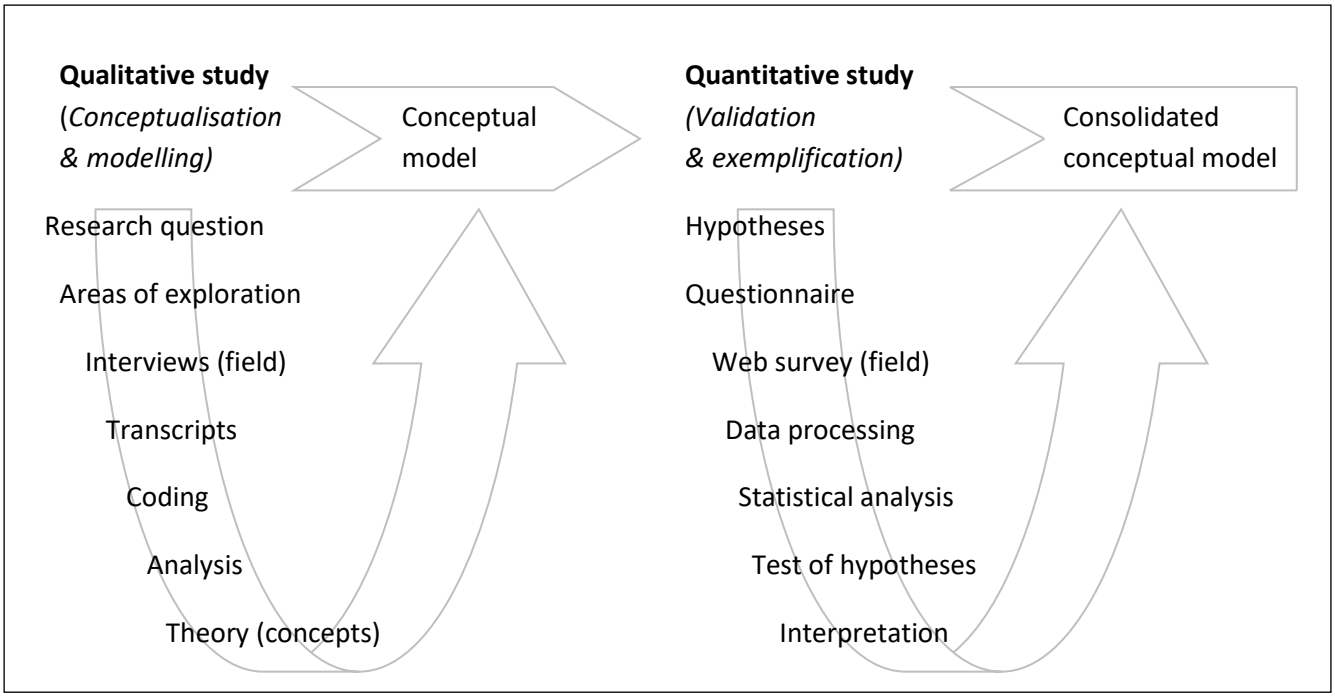


Figure 1 Mixed methods research design

In the qualitative part of the study, experts in data service were interviewed in their role as important intermediaries for data seeking. The interviews were transcribed, coded and analysed. A grounded theory of data seeking behaviour was developed and a conceptual model was constructed on these grounds. For the quantitative part of the study testable hypotheses were drawn from the conceptual model. These hypotheses were employed to design a questionnaire for the quantitative data collection. Data were collected in a web survey among secondary survey data users. With the aim of empirical validation and exemplification of the conceptual model, the collected data were analysed statistically and interpreted against the backdrop of the hypotheses. The analyses were aimed at building a consolidated model of data seeking behaviour.

6. Outline of the Study

This study comprises four chapters. In chapter “B. Theoretical Perspective” the concepts of information behaviour in general and information seeking behaviour in particular are investigated and research that is relevant for the special case of data seeking behaviour is reviewed. The current understandings of the concepts “information behaviour” and “information seeking behaviour” as well as core concepts of information seeking behaviour research are evaluated (B.1). Afterwards, survey data, survey research specifics with regard to possible data seeking behaviours and practices are described on the grounds of past and recent research (B.2). The theoretical chapter ends with specific theoretical assumptions that form areas of exploration for development of a grounded theory of survey data seeking behaviour (B.3).

In chapter “C. Qualitative Study” the methodology, data collection, analysis, and results of the qualitative study are laid out in detail. The chapter starts with a detailed description of the research design of the qualitative study by introducing the chosen *constructivist grounded theory approach* (C.1).The qualitative data collection, the sampling, the coding, and memo-writing are detailed next (C.2). This includes a description of the interview guide, an account of the field phase, the initial sampling and theoretical sampling as well as the coding and analysis using *constant comparative method*. The emerging results are discussed and hypotheses that inform the development of the quantitative instrument are established

(C.3). Subchapter C.3 ends with an account of the validity check. C.4 provides a short summary of the qualitative study.

Chapter “D. Quantitative Study” describes the conduct and outcomes of the quantitative study. First, the methodology and research design are described (D.1). The development of the quantitative instrument, the web questionnaire, is described in detail afterwards (D.2). The data collection is detailed in subchapter D.3. This includes a description of the pre-test and the amendments that had to be made thereupon as well as information on the sampling procedure. The data processing is also described in D.3, followed by the description of the sample in D.4. Subchapter D.5 describes the index and scale development in preparation for the analysis. In subchapter D.6, the statistical data analysis and the results are described in detail. Subchapter D.7 presents the findings. This is done with reference to the theory and model as well as the hypotheses that were established in the qualitative study.

The thesis concludes with chapter “E. Discussion of Results”. Corresponding to the research question, the final chapter gives account of the findings on the information seeking behaviour of survey data users (E.1). In subchapter E.2, the theory and model of the information seeking behaviour of survey data users are presented. To this end, a consolidated model of data users’ information seeking behaviour is depicted. Furthermore – owing to application-oriented research tradition in Library and Information Science – practical recommendations for the design of research data infrastructure are drawn from the results (E.3). Finally, the research contribution is highlighted and prospects for further research are presented (E.4).

B. Theoretical Perspective

The main purpose of this chapter is to lay out which theoretical and empirical work was taken into account in order to prepare the development of theory in the qualitative study. The development of a grounded theory of survey data seeking behaviour is based on relevant theoretical and empirical work that is presented in this chapter. The specific theoretical perspective and preconceptions are brought forward in the following paragraphs, laying the groundwork for the interviews with experts in data service. It is important to note here that only research that had been published before the conduct and analysis of the qualitative interviews could be considered in identifying the areas of exploration that were needed before entering the field. In some cases, the chapter also refers to later work (2017 and after). This is done either to confirm interpretations based on earlier work or to stress the relevance of specific interpretations from an ex-post point of view. Models of data seeking that have been developed and published after the present study had been conducted are included in this chapter only with regard to their confirmation of previous research (Yoon 2017; Yoon and Kim 2017) or are elaborated on in chapter E. Discussion of Results (Gregory, Cousijn, et al. 2019; Gregory, Groth, et al. 2019).

The chapter starts out by declaring the underlying theoretical viewpoint (social constructivism) of this investigation, followed by a short overview of theoretical groundwork in information seeking behaviour that the present study intends to build on. In defining it as goal-oriented problem solving, the present study adopts an understanding of the concept of information seeking that (1) investigates the seeker in terms of their individual characteristics as well as their context; (2) can be analysed in stages and cycles but also in patterns; (3) is purposive, because it considers the seeker's situation as problematic; (4) and encounters barriers.

The second part of this chapter introduces specifics of survey data, survey data infrastructure (data archives) and the survey research process. In reviewing the sparse research on survey data related information behaviour, possible characteristics, practices, purposes, needs, and barriers in survey data seeking are presented. By recurring to more recent research, the supposedly most important context factors in survey data seeking are introduced: the role of documentation, the role of intermediaries, and the role of

information technology. The chapter ends by unfolding areas of exploration for the development of an interview guide for the qualitative study.

1. Studying Information Seeking Behaviour

1.1 Information Behaviour from the Social Constructivist Perspective

This study follows the tradition of a social science perspective in information behaviour research. Within this tradition it leans towards social constructivist approaches of investigating information seeking. Other terms used to denote this research paradigm are collectivism (Talja, Tuominen, and Savolainen 2005) or interpretivist research paradigm (Case and Given 2016; cf. Creswell 2013, 2014). Social constructivism can be understood as one major research paradigm in LIS alongside (cognitive) constructivism and constructionism (according to the terminology used by Talja et al. 2005, who also prefer the term collectivism instead of social constructivism) or, when perceived as the social approach in general, alongside the cognitive approach or multifaceted approaches (Pettigrew et al. 2001). It has also been described as one of the interpretivist research paradigms as opposed to objectivist paradigms (Case and Given 2016). The underlying premise of the social constructivist viewpoint is that "both cognitive processes and the social milieu are important in knowledge formation." (Talja, Tuominen, and Savolainen 2005, 85) More generally, it is assumed that individuals are constructing meaning of situations both subjectively and interactively against the backdrop of their life settings (Creswell 2014). The social constructivist paradigm is, for instance, associated with the socio-cognitive viewpoint and the domain analytic approach introduced by Birger Hjørland and Hanne Albrechtsen (Talja, Tuominen, and Savolainen 2005, 81), with activity theory based on the work of Lev Vygotsky and Alexei Leont'ev as well as sense-making introduced by Brenda Dervin (Case and Given 2016). All of these approaches are considered with *context*, which is a central concept in information behaviour research in general (Agarwal 2017; Case and Given 2016) as well as for the present study. A current definition of information behaviour research as it is carried out here is given by Charles Cole:

> "Information behaviour research looks at the user in a deeper way, below the keywords the user types into the engine's search box. Information behaviour research contextualizes the user by observing and analyzing non-purposive aspects of information seeking motivated by the user's psychology, the cognitive processes whereby users incorporate new information into their prior knowledge to form new or modified knowledge, and the user's sociology; i.e., his or her position in a social group." (Cole 2013)

This definition includes cognitive as well as social aspects, which aligns with a socio-cognitive or constructivist viewpoint.

Another discussion regarding research paradigms and perspectives in information behaviour research surrounds the question whether the activities studied in constructivist investigations can reasonably be called information *behaviour* or should rather be denoted differently, for example, information *practices*, which has been subject to discussion in the field (Savolainen 2007). The notion of *information behaviour* became more and more popular and its widespread use in articles, book titles and curricula (Pettigrew et al. 2001) is conclusive prove for the fact that it is indeed "a term whose time has come" (Case 2012, 91). However, there are several authors who have been criticizing the use of this term for various reasons, their strongest argument probably being the association with a behaviourist research paradigm (Pettigrew et al. 2001). Reijo Savolainen, in particular, ascribes the study of information behaviour to researchers who hold a cognitive viewpoint on how people "deal with information" (Savolainen 2007, 109). He contrasts them with those who hold a social constructionist stance and explains why their research is not about behaviour, but rather information practice (Savolainen 2007). Drawing on "definitions of practice developed in the field of organization science" (Savolainen 2007, 120) Savolainen characterises practice as (1) including repeated and regular actions, (2) embedded in context, and (3) engagement of members of a community in recurrent action. The adoption of the concept of information practice by researchers with social science background came about with a general trend in the social science disciplines that has become known as the *practice turn* (Palmer and Cragin 2008; Rivera and Cox 2014). Savolainen sees information behaviour and information practice as two "umbrella concepts drawing on different discourses that provide a broader context for information studies and suggesting different approaches to metatheoretical and methodological issues" (Savolainen 2007, 109). Indeed, the term information practice is used

by a range of authors who apply social approaches to the study of information seeking and uses (e.g. Caidi, Allard, and Quirke 2010; McKenzie 2003, 2006; Widén-Wulff 2007), but information behaviour is the dominating notion in the discourse,[7] even in studies that obviously follow interpretive approaches (e.g. Mishra, Allen, and Pearman 2015). However, even though the term information practice did not succeed in closing up to information behaviour, it is fair to say that the introduction of practice theory into the discussion about theories and methodology helped raising awareness for the importance of environments or contexts (Talja 2005), and it contributed to the modelling of information activities, in particular with regard to discursive approaches (Talja and McKenzie 2007). A practice view on information activities is also interesting with regard to questions of *information sharing* (Savolainen 2007), a concept that is of relevance for secondary researchers looking for data. Furthermore, in particular with regard to scholars' information practices, who act within a context of disciplinary specifics and scholarly communities (cf. Palmer and Cragin 2008), there is an affinity with domain analysis according to Birger Hjørland and Hanne Albrechtsen , which is a relevant methodology for the present study. Practice based approaches in general have also been associated with Activity Theory, ethnomethodology, and Actor Network Theory (Rivera and Cox 2014). The present study investigates a special case of information seeking from a social constructivist point of view, which considers social context a most relevant factor. In that regard, the insights from practice research in information seeking may prove useful in the case that is under study here. However, the author holds the view that insights from studies of information behaviour are not to be discarded in their relevance; this is because researchers of information behaviour are neither to be confused with behaviourist researchers (cf. Wilson 2009) nor to be reduced to researchers with a cognitive viewpoint. Or as Tom Wilson put it, "'[h]uman behaviour' is about how people act in the world, and it is well understood that a person's actions have both cognitive and social dimensions." (Wilson 2009 n. pag.) From an analysis of the literature one may gain the impression that behaviour studies are more focused on needs and motives, while practice studies emphasise social and cultural factors (Savolainen 2007). However, it remains

[7] This is apparent, for instance, in the proceedings from the Information Seeking in Context (ISIC) conference. The proceedings from 2018 (http://www.informationr.net/ir/23-4/isic2018/isic2018.html and http://www.informationr.net/ir/24-1/isic2018/isic2018.html, accessed November 11, 2020) show a clear dominance of the term "information behaviour", even though what is studied might actually be information practices.

doubtful whether the debate around the umbrella concepts is in fact mirrored in a substantial divergence of preferred concepts, literature, and methods in information (seeking) research (Case and Given 2016). In any case, it is reasonable to conduct research on the premise that the two concepts are rather complementing than excluding each other.

The influence of the social constructivist perspective as a theoretical lens in this study influences (1) the research questions and (2) the setting of goals as well as (3) the choice of methods and (4) the data analysis (cf. Creswell 2014).

Ad (1), with regard to the main research question, the social constructivist perspective leads to the disciplinary focus on users of survey data, because field-specific schools of thought, norms, and research practices are assumed to be determinant factors of context in data seeking.

Ad (2), the goal of the enquiry is social constructivist in that it aims at improving practice through interpreting and understanding constructed meanings of data seeking behaviour in order to develop a behavioural model as well as concrete recommendation for infrastructure development, thus by choosing a mediating way between theoretical and practical research (cf. Lincoln, Lynham, and Guba 2011).

Ad (3), the social constructivist perspective also influences the choice of methods in this study in starting from qualitative interviews that are supposed to enable the participants to construct subjective meanings and the researcher to interpret them hermeneutically (cf. Lincoln, Lynham, and Guba 2011). For theory development, intermediaries from secondary data services are interviewed instead of talking directly to the users. This approach is also influenced by social constructivism in that the events that the participants will talk about are subjective experiences of interaction, rendered by the participants' implicit meanings (Charmaz 2014).

Ad (4), with regard to data gathering and analysis, it is assumed that the interviews with knowledgeable experts result in more reliable categories drawn from their condensed experiences – a fact that is advantageous in the development of the second part of the study, the web survey with secondary users of survey data. The subjective meanings of the interviewees are viewed here as a meta-perspective on the collective construction of reality.

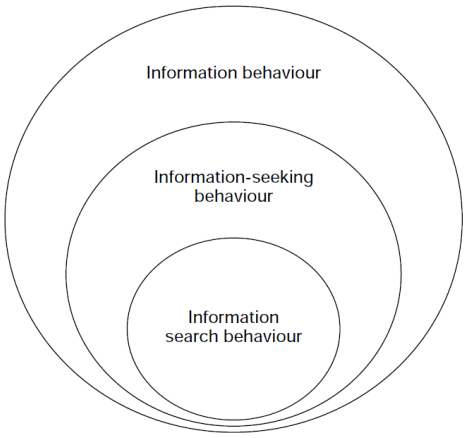
Most importantly, for the analysis and the grounded theory development the interview data is understood in the sense of “interpretive renderings of reality, not objective reportings of it” (Charmaz 2005, 510).

The theoretical perspective or lens of social constructivism provides the interpretive framework for this study; it does not serve as a testable theory. The theoretical lens of social constructivism is seen as a research paradigm in the sense of a “basic set of beliefs that guide action” (Guba 1990, 17). However, it leads the process of *constructing* such a theory inductively from the data gathered in the qualitative part of the study. The view that grounded theory is rather *constructed from* than *discovered in* data is adapted from Kathy Charmaz whose interpretive framework can be seen as social constructivist as well (Charmaz 2005). Charmaz holds the view “that any theoretical rendering offers an *interpretive* portrayal of the studied world, not an exact picture of it.” (Charmaz 2014, 17). Following an "inductive logic of research" (Creswell 2014, 65), existing concepts and theories from LIS research, in particular those that can be associated with the constructivist paradigm, influence the design of the interview, the data analysis, and the theory building.

Various authors coming from all metatheoretical traditions have engaged in modelling the concepts of *information behaviour* as well as *information seeking* or *searching* and further models are developed as research advances. The information seeking models in particular are numerous (Saracevic 2009). An often cited illustration of the research field that provides a rough classification and analytical distinction of models of information behaviour, seeking, and searching has been published by Tom Wilson in his nested model of the research areas (see Figure 2).

With his nested model, Wilson puts the three concepts of *information behaviour*, *seeking*, and *searching* in a hierarchical relationship. Particularly, he suggests that information seeking on the one hand and information searching on the other hand are different phenomena within information behaviour that are to be analysed complementary but not interchangeably (Wilson 1999). Wilson also gave distinct definitions for all three research areas that have been adopted widely. In 1999 he defined the all-encompassing information behaviour as “those activities a person may engage in when identifying his or her own needs

for information, searching for such information in any way and using or transferring that information" (Wilson 1999, 249).



A nested model of the information seeking and information searching research areas

Figure 2 Wilson's nested model of research areas (Wilson 1999, 263)

In the following year, Wilson developed his understanding of the concept further to a definition that is broadly cited to this day (K. E. Fisher, Erdelez, and McKechnie 2005; Lloyd and Olsson 2017; Scheibe, Fietkiewicz, and Stock 2016; Wijetunge 2018): "Information Behaviour is the totality of human behavior [sic] in relation to sources and channels of information, including both active and passive information seeking, and information use." (Wilson 2000, 49) A more recent, quite similar definition is given by Donald Case and Lisa Given: „*Information behavior* [...] encompasses information seeking as well as the totality of other *unintentional* or *serendipitous* behaviors (such as glimpsing or encountering information), as well as purposive behaviors that do not involve seeking, such as actively *avoiding* information." (Case and Given 2016, 6) Both definitions revolve around *information seeking*, the core concept in the field that has been studied widely and been described in many models. According to Wilson's nested model, information seeking involves *information searching*, the latter being restricted to behaviour that refers to user interaction with information systems (notably computer-based information systems) (Savolainen 2016; Wilson 1999). This understanding of information searching (or information search behaviour) positions this concept close to the realm of information retrieval research (Courtright 2007; Saracevic 2009; Savolainen 2016). For the present study, however, the

broader concept of information seeking is the phenomenon in question. Information seeking is one of the early concepts that have been studied in the field of information science and also in other disciplines such as psychology, sociology, and political science (Saracevic 2009).

1.2 Information Seeking Behaviour

Wilson defines information seeking as "the purposive seeking for information as a consequence of a need to satisfy some goal" (Wilson 2000, 49). This definition has been widely adopted in the literature, for example by Eszter Hargittai and Amanda Hinnant, who favour a social approach to studying information behaviour: "Information seeking is now just purposive information seeking as information-seeking research has traditionally examined the problem situation of the user, and the purposive information seeking done to solve the problem" (Hargittai and Hinnant 2006, 57). Which activities are encompassed by information seeking differs from author to author (e.g. Borgman 2007), resulting in a broad range of elements that have been treated and analysed as seeking behaviour or practice of seeking in past and present (Courtright 2007). However, the relevance of goals, purpose, and problems or *problematic situations* (Wersig and Windel 1985) is stressed by a majority of authors (cf. Saracevic 2009), as is a relatively broad understanding of the concept in that it includes active, passive, directed, and undirected behaviour (Courtright 2007; pace Case 2012; Savolainen 2016). Furthermore, many models imply a procedural understanding of information seeking, e.g. Carol Kuhlthau's *Information Search Process* (ISP) (cf. Saracevic 2009), whereas others illustrate behavioural patterns that don't necessarily occur sequentially, e.g. David Ellis' *Characteristics of information patterns* (Wilson 1999). Some researchers have proposed to describe information seeking more iteratively or in cycles rather than in stages (Blandford and Attfield 2010; Marchionini 1995; Pontis et al. 2017). Reijo Savolainen concludes that while the approach to study information seeking in cycles has gained relevance more recently and especially in the context of online environments, these models should rather be used complementary than opposed to models that focus on stages of information seeking (Savolainen 2018). The present study follows this suggestion by considering that information seeking can occur in patterns, stages, and cycles.

In particular, the results of studies by David Ellis and colleagues (Ellis 1989; Ellis, Cox, and Hall 1993) have been incorporated in many other studies and examinations (e.g. Azama and Fattahi 2011; Bronstein 2007; Choo, Detlor, and Turnbull 2000; Ge 2010; Meho and Tibbo

2003). Today, about thirty years afterwards, researchers still employ their characteristics of information patterns and seek to test and adapt them according to their own research questions (e.g. Fitzgerald 2018; Weigl et al. 2017). Their prevalence in the field make the ideas by Ellis et al. particularly interesting to follow up in the present study.

Ellis' (1989) analysis of behavioural patterns in seeking resulted in the six characteristics of information seeking behaviour in the social sciences: *starting*, *chaining*, *browsing*, *differentiating*, *monitoring*, *and extracting*. Ellis wanted the six categories to be viewed as features of a model that "represent the major generic characteristics of the social scientists' individual information seeking patterns" (Ellis 1989, 178). He explained the characteristics as follows:

- *Starting*: activities characteristic of the initial search for information;
- *Chaining*: following chains of citations or other forms of referential connection between material;
- *Browsing*: semi-directed searching in an area of potential interest;
- *Differentiating*: using differences between sources as filters on the nature and quality of the material examined;
- *Monitoring*: maintaining awareness of developments in a field through the monitoring of particular sources;
- *Extracting*: systematically working through a particular source to locate material of interest. (Ellis 1989, 178)

In explaining the features of his model further, Ellis delivers some results concerning the actual information seeking behaviour of social scientists. For example, when *starting* to work on a new topic or in a new area, they prominently employed the use of personal (or informal) contacts – a pattern that before Ellis many other researchers in social science information behaviour had discovered (Ellis 1989, 179). Informal channels of information (e.g., *consulting a colleague*) have been primary information sources for researchers in all kinds of fields (Case and Given 2016; Cronin 1982). For example, already in 1961, the American Psychological Association (APA) found that psychologists relied heavily on informal information channels such as conventions as well as formal information such as published articles (Paisley 1965). Authors of articles found it difficult to find and access current material and described difficulties with existing indexing services. A couple of years later, in 1967, a large scale investigation on information needs of social scientists called "Information

Requirements of the Social Sciences" (INFROSS) yielded similar results. The studied researchers and practitioners from the fields of anthropology, economics, education, political science, psychology, and sociology relied heavily on informal channels of information (such as colleagues pointing them to relevant publications) and were rather unsatisfied with formal information channels (such as library catalogues) (Hunsucker 2007; Line 1971). Later on, there have been further studies that emphasise the importance of informal channels of information and rather low usage of formal channels of information such as library catalogues and indexing services (Folster 1995; Satish 1994). The importance of informal channels of information in information seeking behaviour has led to the widely adopted concept of "invisible colleges" (Allen 1969, 4). The *invisible college*, understood as the personal context of a researcher, has been a significant factor even in early conceptualisations of information behaviour (Cronin 1982; e.g. Ford 1977).

Also, researchers tended to look for introductory works, key references and key authors to begin their work (Ellis 1989). The interviewed psychologists in particular, indicated to make use of reviews and review articles when *starting*. Only some interviewees also named formal channels of information as relevant for *starting*, e.g. bibliographies, abstracts, indexes, library catalogues (Ellis 1989). Ellis found that *starting* activities were often applied with the goal to find some basis for *chaining*. He described this second characteristic of information seeking as taking two forms: *backward chaining* (the traditional way of identifying sources by following citations in a known relevant source) and *forward chaining* (the relatively new way of identifying sources that cite a known relevant source). Backward chaining was employed by all researchers interviewed by Ellis, while forward chaining was rarely used – a result that has to be viewed in context of the time of survey, when citation indexes were not that commonly used.

Going on with Ellis' results, *browsing* activities were employed by many of the participants in the study (Ellis 1989). Typically, the surveyed social scientists browsed by scanning content pages of journals, checking periodicals, and browsing along library shelves. All interviewed individuals also used ways of *differentiating* material, that is to say, they filtered information sources according to their potential usefulness. Ellis identified three most significant criteria for this information seeking feature: (1) the substantive topic of study; (2) the approach or perspective adopted; and (3) the quality, level, or type of treatment. In terms of *monitoring*,

the interviewed social scientists employed different strategies: use of informal contacts; use of monitoring services; use of research directories; use of publishers' catalogues; and reading journals or newspapers. Finally "one of the most directed and focussed of information seeking activities" (Ellis 1989, 198) that the studied researchers employed was *extracting*, for example from "a run of a periodical, a set of conference proceedings, a series of monographs, the contents of an archive, a collection of publishers' catalogues, or bibliographies, indexes, or abstracts" (Ellis 1989, 198). Since it is a thorough and time-consuming activity, the identification of sources suitable for *extracting* is critical. As Ellis found out, suitable sources are either recommended by colleagues or supervisors or have a correspondent standing in the field.

In 1993, David Ellis and colleagues published their results from a follow up study that was aimed at comparing the social scientists' information seeking characteristics with those identified in physicists and chemists. To ensure comparability, the authors chose a similar methodological approach to the one applied in Ellis' first study (Ellis, Cox, and Hall 1993). They concluded that the analyses rendered no remarkable differences between the information seeking patterns of social scientists and physicists. The six characteristics of information seeking of social scientists identified some years earlier were applicable to the physicists. In the case of the chemists, the authors found two further characteristics, which they labelled *verifying* and *ending*:

- *Verifying*: activities associated with checking the accuracy of information;
- *Ending*: activities characteristic of information seeking at the end of a topic or project, for example, during the preparation of papers for publication. (Ellis, Cox, and Hall 1993, 359)

In hindsight, Ellis et al. noted, that some of the social scientists studied earlier had employed patterns in the categories of *chaining* and *starting* that could equally be subsumed under one of the two new characteristics. Consequently, the authors concluded in line with earlier comparative work by other researchers that "[…] there are not major differences in the information seeking activities of social scientists and scientists although there are differences of emphasis." (Ellis, Cox, and Hall 1993, 366) Going further, a study of information seeking patterns of engineers and research scientists in an industrial

environment by Ellis and Haugan (1997) found that the behavioural characteristics identified were similar to those found in the previous studies of academics.

The fact that both studies followed qualitative approaches makes it of course plausible that the identified characteristics be supplemented or rearranged in later investigations. Also, technical developments and the establishment of new theoretic and methodological approaches in a field over time should render corresponding developments in information seeking behaviour. To illustrate this point, it seems worthwhile to investigate a little further in the *verifying* characteristic. Ellis et al. bring up this feature by explaining how "most of the chemists indicated that they were aware of the possibility of errors, particularly typographical errors, occurring in their information. Errors in numerical data were the most commonly cited – although other errors, for example in citations, nuclear magnetic resonance (NMR) assignments and equations were also mentioned." (Ellis, Cox, and Hall 1993, 364–65) That activities like these are critical in a scientist's research process is obvious, because of the impact that numerical errors can have compared to, for example orthographic errors. From today's perspective we can add that this is no less true for the research process of social scientists, at least if their work is empirical. One reason why *verifying* did not emerge as a clear characteristic of social scientists' information seeking in Ellis' 1989 study may be that, at the time, empirical social research did not yet have the dominant standing that it has today and hence was not practiced by the researchers in Ellis' sampling. However, for the present investigation of survey data users, their employment of *verifying* activities with regard to survey data was expected to be of particular interest.

As for the other characteristics of information seeking identified by Ellis, it is of interest to what extent they occur in data seeking. For example, *chaining* between datasets cannot directly employ citations, since there are no citations between datasets. There is, however, the conceivable possibility of *forward chaining* from literature to a cited dataset, and of *backward chaining* from the record of a dataset in a data catalogue to publications that have used (and cited) this data. These characteristics are revisited in subchapter B.2.1.2 in context of the survey data research process.

One particular study that built on Ellis' work is the investigation on information seeking behaviour of social scientists studying stateless nations conducted by Lokman Meho and

Helen Tibbo (2003). Meho and Tibbo aimed at updating Ellis' work by extending the sampled population beyond the borders of a single university and by considering the World Wide Web as a critical influence. By choosing to interview researchers who study stateless nations, the authors arrived at a studied population that represented a large variety of disciplinary affiliations; interviewees came from anthropology, area studies, communication, economics, education, geography, history, political science, psychology, public administration, sociology, and women's studies (Meho and Tibbo 2003). In their analysis, the authors found clear evidence for Ellis' (1989) six characteristics of information seeking, but they also found it necessary to complement the model by adding the characteristics *accessing*, *networking*, *verifying*, and *information managing*. Problems with information *access* seemed to be an issue for a majority of the studied researchers. As for the *verifying* characteristic, Meho and Tibbo did consider that it had already been added by Ellis, but rather with regard to physical scientists and engineers. In the context of the newer study by Meho and Tibbo, this characteristic gained new impact through the availability of online information. Participants of the study indicated that they found interesting material on the Internet, but encountered difficulties in "verifying its legitimacy or its source" (Meho and Tibbo 2003, 582).

Another effect resulting from the existence of the Internet was found in the characteristic of *networking*, that the authors interpret with reference to the concept of *invisible colleges* (see above) (Meho and Tibbo 2003, 583). For Meho and Tibbo, the concept seems to have become even more important through advanced technology. It is remarkable in this context, that the most important means of keeping up-to-date indicated by the interviewees were subscriptions to listservs (indicated by 24, followed by journal subscriptions, indicated by 20). Finally, *information managing* tasks were indicated "repeatedly" by the stateless nations researchers and the authors deemed these activities highly significant for information retrieval given that the studied researchers also indicated to rely heavily on their personal collections. In the end, Meho and Tibbo arrived at a model of information seeking behaviour of academic social scientists that arranged the identified characteristics around four stages: *searching*, *accessing*, *processing*, and *ending*.

Apart from its role as a stage in Meho and Tibbo's model of information seeking the activity of *searching* deserves to be specified further at this point. As Carol Tenopir and Donald W. King (2008) have shown empirically, the means of finding relevant journal articles employed

by academics have changed substantially over the last decades (Tenopir and King 2008). Due to the development and spread of online databases and web search engines, the proportion of *automated searching* has increased from 0.7 % in 1977 to 23.1 % in 2005 (Tenopir and King 2008). Accordingly, when David Ellis conducted his research back in the 1980s, in pre-Internet times, *searching* in a formal sense, like keyword or bibliographic searching, using library subject catalogues or online databases, occurred in social scientists' information seeking behaviour mainly when they were starting their research career or familiarizing with a new research area (Ellis 1989). Consequently, *searching* did not end up as a characteristic of its own in Ellis' model, but was considered as a way of *starting*, alongside identifying *starter references* or *reviews and review articles* (Ellis 1989). Particularly online searching was not employed much back then, and if online databases were used, the satisfaction with the results was rather low (Ellis 1989). Along with an increase in electronic publishing, online searching as a means to locate scientific information has increased as well (Tenopir and King 2008). Nowadays, online searching is a natural way of looking for information (Athukorala et al. 2014) that can even exceed information seeking via personal contacts (Tenopir and King 2008). Unfortunately, the applicability of these findings for the case of data seeking and usage is limited. In the case of data seeking, the role of intermediaries may still be greater. In their 2009 study of social scientists' perception of data documentation quality, Jinfang Niu and Margaret Hedstrom found that:

> "Some users of this kind of data are willing to take on the additional burdens associated with using large complex datasets. They think it is natural that documentation does not provide everything. It is part of their research process to interact with other researchers for secondary research use." (J. Niu and Hedstrom 2009, 129)

In general, it remains to be seen whether results of seminal research such as Ellis' studies are transferrable to data seeking. *Purposes*, *patterns* of seeking, the role of *information technology* as well as relevance of *intermediaries* are factors in information seeking that is investigated further with regard to survey data seeking in subchapter B.2.3. Other important influences that prevail as concepts in information seeking behaviour research are *context* and *domain* as well as *individual* factors in the information seeker. Presumably, these

concepts will be all the more relevant for the case of seeking research data. These factors will therefore be reviewed in more depth in the following section.

1.3 Individual, Context and Domain in Information Seeking Behaviour

By adopting a social constructivist viewpoint, *context* is seen as a central concept for the understanding of information activities, and of information seeking in particular (cf. Pettigrew et al. 2001). Context is one of the key concepts in information behaviour research (Agarwal 2017). The preoccupation with this concept can be traced back to early studies of information behaviour, for example to the work of William Paisley (1968), G. Ford (1977) or Robert Taylor (1991). Context has been and still is of particular importance in studies of information seeking, even though it is a broad concept that makes it difficult to compare context-related studies (Saracevic 2009). The notion of *context* is by all means not used consistently in the field (Agarwal 2017). Some researchers use different notions to refer to contextual factors, for example: *situation* (Brenda Dervin 2003; pace Courtright 2007); *frame of reference*; *setting*; *environment*; *information world*; *life world*; *information ground* (Courtright 2007).

Tom Wilson, for instance, emphasises the importance of context in information seeking. In his view, the information seeker is subject to their needs and surrounded by demanding roles and environments. A very clear and elaborated incorporation of contextual factors into the concept of information seeking can be identified in his model of information-seeking paths that he demonstrated in his seminal article from 1981 (Wilson 1981). Figure 3 shows a slightly revised version of this model that Wilson published and renamed *The information user and the universe of knowledge* in 2005.

In striving to portray the possible surroundings of the information seeker and the complex interactions of these worlds (Wilson 1981), this model depicts (1) the seeker's context, (2) the employed system, and (3) information resources (Wilson 2005). In that regard this early model of information-seeking already resembles more current views of social context, situation and environment, for example defined as incorporating the whole of physical, social and technological factors (cf. Bates 2010).

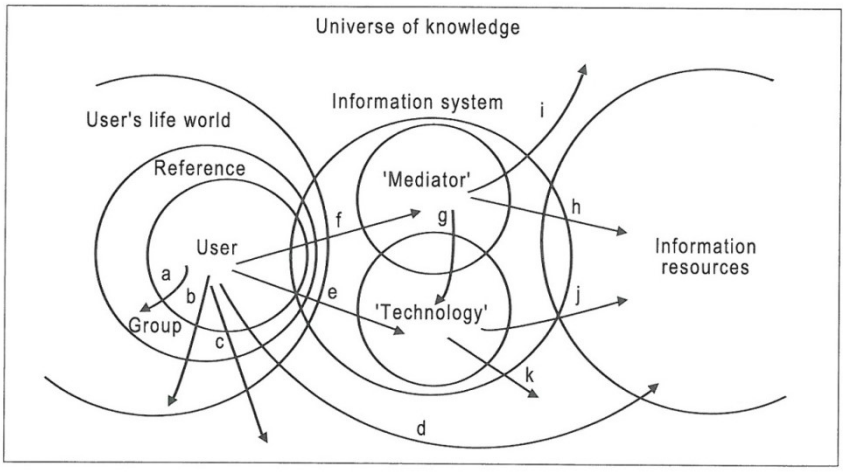


Figure 3: The information user and the universe of knowledge (Wilson 2005, 32)

While in this model Wilson depicts the user as surrounded by their life-world "defined as the totality of experiences centred upon the individual" (Wilson 1981, 6), he then moves on and introduces a more detailed view on this life-world in his model of *Factors influencing needs and information-seeking behaviour* (later renamed *Information need and seeking*, see Figure 4). In this model, Wilson takes basic human *needs* (physiological, affective, or cognitive) as a starting point and places them "at the root of motivation towards information-seeking behaviour" (Wilson 1981, 9), while emphasising that "these needs arise out of the roles an individual fills in social life" (Wilson 1981, 9).

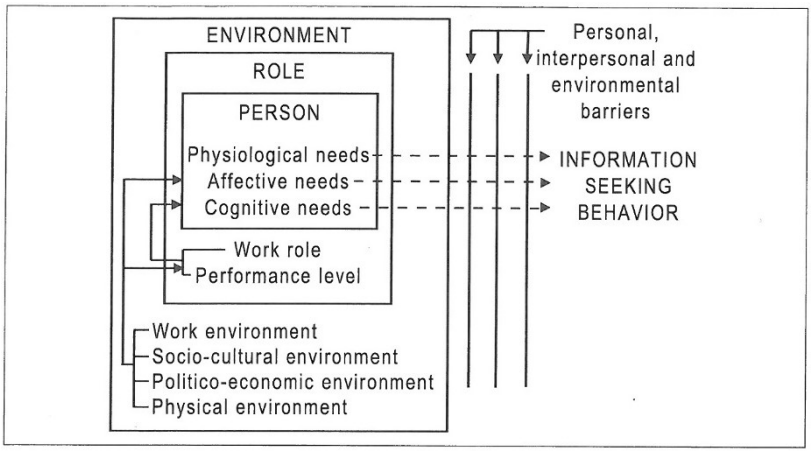


Figure 4: Information need and seeking (Wilson 2005, 33)

In addition to the contextual factors, the model also considers possible intervening *barriers* to information seeking. Barriers of different origin provide an important analysable concept with regard to context in information seeking (Brown 1991). For the case of searching e-journals, Xuemei Ge (2010) identified several obstacles: lack of availability; lack of accessibility; usability issues; uneven source quality; disciplinary and research topic constraints; perceived ease of use; lack of awareness; personal constraints (Ge 2010). Referring to Wilson's model, Mary Brown synthesised: "The *self*, the *role*, and the *environment* form the foundations of the *context* of information-seeking behavior." (Brown 1991, 10) The influence of different life-worlds on communication (or information activities) is also underlying the domain-analytic approach as it is advocated by Birger Hjørland, who explains that "[t]his view changes the focus of IS from individuals (or computers) to the social, cultural, and scientific world" (Hjørland 2002, 258).

In his recent book on context in information behaviour, Naresh Kumar Agarwal proposes another analytical layer for interpretivist studies (Agarwal 2017). Agarwal differentiates between individual, shared, and stereotyped views of context by arguing that positivist research designs may claim to analyse the individual or personal perspective (the "self" or the "role"), when in reality they tend to look at information behaviour from a stereotyped or out-group point of view. Agarwal's definition of context caters to these different viewpoints and illustrates the many possible understandings of contexts of information behaviour research:

> "The context of an actor's information behaviour consists of elements such as environment, task, actor-source relationship, time, etc. that are relevant to the behaviour during the course of interaction and vary based on magnitude, dynamism, patterns and combinations, and that appear differently to the actor than to others, who make an in-group/out-group differentiation of these elements depending on their individual and shared identities." (Agarwal 2017)

The present study considers cognitive and social context factors and acknowledges that context can be viewed and analysed from different viewpoints on a spectrum from in-group to out-group perspectives. As Agarwal has shown in his literature review and also considered in his proposed definition of context, there have been multiple attempts to collect particular factors that have been identified as contextual in research. For the present study, all mentioned collections of contextual factors are of interest in that they can support the

coding of the data in the qualitative study and thus help develop a grounded theory of data seeking. It is therefore neither necessary nor possible to define factors of context in advance. Instead, this study starts by recognizing context as a major concept to focus on in the investigation and therefore operates with a very broad definition of context, such as "the place where meaning is socially constructed" (Tabak 2014, 2225).

Investigating information behaviour from a social viewpoint, focusing on people in their contexts is usually done by investigating groups, for example by studying people in particular work, organisational, or social settings (Case and Given 2016; X. Niu et al. 2010). The group-oriented approach of studying information behaviour originated in studies by occupation or academic *discipline*, (Bawden and Robinson 2013; and already Ford 1977) and until today the investigation of information seeking in academia remains one of the main areas of interest (Borgman 2007; K. Fisher and Julien 2009; Herman 2004a, 2004b; Palmer and Cragin 2008). What is commonly meant by the term *discipline* in research on scholarly information practice is, for example, described by Carole Palmer and Melissa Cragin who explain that it is "used to describe and differentiate knowledge, institutional structures, researchers and resources in the working world of scholarship and science" (Palmer and Cragin 2008, 172). Referring to J.T. Klein, they add: "Disciplines […] represent subject areas, tools, procedures, concepts, and theories of stable epistemic communities" (Palmer and Cragin 2008, 173). Discipline-oriented or field-specific investigations have shown that there are indeed differences in information seeking behaviours or practices of researchers from different domains (Talja 2005), for example with regard to the use of electronic journals (cf., for example, Talja and Maula 2003; Tenopir et al. 2010; Tenopir, King, Spencer, et al. 2009). For the special case of data practices, it has also been reported that they "are influenced by researchers' disciplines and subdisciplines […]." (Palmer et al. 2009, 32)

Even though there are other factors that have been identified to be influencing information behaviour of academics in general – for example, age, work role and responsibilities, motivation or purpose (Tenopir, King, Spencer, et al. 2009) – the disciplinary differences have repeatedly shown to have an effect. Bradley M. Hemminger and colleagues hold that even though "at a high-level" strategies of information seeking can be similar across disciplines, differences between fields do occur (Hemminger et al. 2007). For instance, Carole Tenopir and colleagues found in a study of 1,688 US American and Australian

researchers from different disciplines that subject discipline as well as work responsibilities were the strongest factors influencing article seeking and reading, while productivity, age, and purpose accounted for less influence. (Tenopir, King, Spencer, et al. 2009) Many other researchers have studied patterns of scholarly information behaviour in similar or different ways; Jenny Fry, in summarising key findings from this research concludes:

> "That scientific communication is embedded within a context of scholarly tradition and that the forms and technologies of communication are shaped by disciplinary rituals and practices." (Fry 2006, 301)

Considering the amount of studies of this kind, it is obvious that the investigation of researchers' information behaviour has influenced the development of many models, theories and approaches in the field. Most apparently, *domain analysis* is a discipline-oriented approach that has been developed "with the goal of forming holistic understandings of scholarly communities' work and communication practices" (Talja 2005, 123). This approach has been introduced to information science by Birger Hjørland and Hanne Albrechtsen in 1995 and can be seen as one of the most influential ideas related to the so-called *social turn* in the discipline.

With Hjørland and Albrechtsen it can be argued that for the understanding of researchers' information behaviour it is important to consider their disciplinary surroundings such as communication, subjects, paradigms, the function of information systems and structures in their knowledge domains (Hjørland and Albrechtsen 1995). The nature of a specific domain leads to a better understanding and the possibility to make generalizations on information seeking within this domain (Hjørland and Albrechtsen 1995). Hjørland explains the influence of the domain on information searching:

> "When we speak of people's relevance criteria in relation to IR [Information Retrieval], they are […] mainly determined by cultural factors. They may, for example, be determined by trends or 'paradigms' in knowledge domains […]. When searching for literature about a topic […] the relevance criteria are implied by the theory, tradition or 'paradigm' to which the searcher subscribes or belongs." (Hjørland 2005, 339)

For the case of information retrieval, he specifies that activities are determined by "trends and 'paradigms' in knowledge domains" (Hjørland 2005, 339) and that "[r]elevance criteria are socialized into the individual from the academic tradition in which he has been raised" (Hjørland 2005, 339). With regard to data seeking that is in focus of this investigation, the assumptions that Hjørland makes for the relevance of literature are supposed to be even more applicable, because research data characteristics and properties vary across domains to a much greater extent than with regard to literature. These variations are influenced by research paradigms or methodology, not only between disciplines, but also within single fields, for instance in sociology where researchers gather data with several different methods, according to the schools of thought that they belong to, the fellow researchers that they work with, the research paradigms that they follow etc.

A major subject of discussion among domain-analytic researchers is the understanding and operationalisation of the concept *domain* (cf. Fry and Talja 2004, 21). In Birger Hjørland's understanding, a domain is not necessarily a scientific discipline, like the social sciences. This assumption is in line with empiric research, for example by Jenny Fry (2006) who found that in terms of communication patterns researchers from different fields within a particular discipline can have less in common than researchers from different disciplines (Fry 2006). In Hjørland's view, a domain is composed of a "discourse community, being a community in which an ordered and bounded communication process takes place" (Hjørland 2002, 258). Carole Palmer and Melissa Cragin name a few alternatives to the understanding of domains as disciplines, among them areas "where researchers cooperate in sharing data, tools and expertise" (Palmer and Cragin 2008, 178). This particular understanding of a domain is applicable to the present study of secondary researchers who are re-using social science data gathered by others.

Hjørland emphasises that, first and foremost, analyses of researchers' information seeking should be based on "a theory about the information seekers' […] interpretation of the sources and concepts in the field" (Hjørland 2000, 38). As a frame of reference, he refers to the metatheoretical paradigms that a researcher follows in their field, arguing that "[e]pistemological theories are our most general models of how people look at their respective fields" (Hjørland 2000, 38). Hjørland exemplifies this approach in a study of relevance criteria in information retrieval for the field of psychology (Hjørland 2002). He

presents a simplified matrix of relevance/non-relevance criteria for the four epistemological schools of *empiricism*, *rationalism*, *historicism*, and *pragmatism* (see Table 1).

Table 1 Simplified relevance criteria in four epistemological schools (Hjørland 2002, 269)

Empiricism	Rationalism	Historicism	Pragmatism
Relevant: Observations, sense-data. Induction from collections of observational data. Intersubjectively controlled data.	*Relevant:* Pure thinking, logic, mathematical models, computer modeling, systems of axioms, definitions, and theorems.	*Relevant:* Background knowledge about preunderstanding, theories, conceptions, contexts, historical developments, and evolutionary perspectives.	*Relevant:* Information about goals and values and consequences both involving the researcher and the object of research (subject and object).
Nonrelevant: Speculations, knowledge transmitted from authorities. "Book knowledge" ("reading nature, not books"). Data about the observers' assumptions and preunderstanding.	*Low priority* is given to empirical data because such data must be organized in accordance with principles that cannot come from experience.	*Low priority* is given to decontextualized data of which the meanings cannot be interpreted. Intersubjectively controlled data are often seen as trivia.	*Low priority* (or outright suspicion) is given to claimed value free or neutral information. For example, feminist epistemology is suspicious about the neutrality of information produced in a male dominated society.

This approach has been lauded and criticised at the same time, on the one hand as qualifying for a major contribution to information science while on the other hand giving too much priority to epistemology instead of the cultural or social world (Cf. Fry 2006; Fry and Talja 2004). For the present study that is in fact not investigating representatives of a specific discipline (e.g., the social sciences), but rather researchers who work with empirical data, mostly gathered in social science surveys, Hjørland's matrix is indeed very useful when trying to shape the domain of the studied researchers: It is assumed here that the average survey data seeker has an *empiric* focus on research. According to Hjørland's matrix, this researcher on the one hand deems observational, intersubjectively controlled data as relevant. They would on the other hand be less interested in speculations, knowledge transmitted from authorities, book knowledge, assumptions and preunderstanding.

Following Hjørland's understanding of domains as "discourse communities", the community of secondary users of survey data can be thought of as a domain. The supposedly prevalent empiric focus in this community is one feature of the domain that is expected to be of relevance for the secondary users' data seeking behaviour. Other factors such as "education, training, professional development, and […] reputation building" (Fry and Talja 2004, 22) may indeed be influenced by the specific discipline or scientific field that members of the community come from. In the case of secondary researchers looking for survey data, a

certain variety of research fields come into question, for example sociology, political science, psychology, or economics (and all these fields can again be subdivided). It is worth noting that secondary users of survey data do not necessarily have to be academics, but can, for example, also be journalists or private researchers. However, since working with empirical data requires sophisticated skills usually obtained and developed in academic contexts, the majority of users supposedly come from academia.

The present study follows the ideas of domain-analysis insofar as it is assumed here that general knowledge of the domain is an important point of departure for the design of the qualitative part of the study, in particular for the theoretical assumptions that will be presented in this chapter. It is expected here that the domain-analytic view is especially fruitful with regard to data seeking as opposed to literature seeking. Most strikingly, sharing and reuse of data have a long tradition in some disciplines (such as the empirical social sciences), while other disciplines (like ecology, cf. Zimmerman 2007) have only just begun to employ secondary data use (cf. Zimmerman 2007). In general it is assumed here that disciplinary differences are all the more important when it comes to data reuse as opposed to literature use. Recent research supports this assumption (Faniel, Kriesberg, and Yakel 2012; Palmer et al. 2009; Tenopir et al. 2011).

2. Studying Data Seeking Behaviour

When studying data seeking behaviour, the rich knowledge from general research in information seeking behaviour as it has been presented in excerpts and exemplary here, provides a very useful starting point. In particular, the conceptualisations of context, domain, purposes, patterns of seeking, barriers, the role of intermediaries as well as information technology are interesting to follow up in the present context. The relatively sparse research on data seeking behaviour suggests specifics in these areas. Additionally, data-specific influencing factors such as data documentation and data literacy of the researcher who is looking for data seem to be of relevance, as the existing research suggests.

In the following paragraphs, all these aspects will be addressed. Following the domain analytic approach, the specifics of survey data and survey data archives are discussed first. Secondly, the survey research process is described and assumptions are made how this

domain-specific process influences possible patterns or practices of data seeking. This section is followed by an estimation of needs and purposes as well as barriers that might occur when looking for data. The section ends with an overview of the data-specific influencing factors. Special attention is given to the importance of data documentation, the role of intermediaries and of information technology.

2.1 Context and Domain of Survey Research

The domain of empirical social research in particular is pivotally concerned with experience or observation of human behaviour (Punch 2013). It is about collecting and analysing observable information, recorded as data (Punch 2013), and only these data are of relevance for empiricists. However, research data used in the social sciences are manifold, and surveys are only one of many sources of social science data. Other main sources of data are: experiments; public records; historical documents; statistical yearbooks; and direct field observation (Lewis-Beck 2004a). Additionally, social researchers increasingly use data that have not been gathered for research purposes, such as transactional data that may arise from electronic payment or from internet activity (Quandt and Mauer 2012). According to the source and method of gathering, social science research data appear in different forms, including: Audio/ video recordings; protocols, transcripts; diaries, biographies, narratives; statistics, including official statistics and numeric results from surveys; transactional data; tables (e.g. SPSS files, Excel files); output files from coding software (Herb 2015). Of particular relevance for empirical social research are so-called micro-data that contain information on each studied unit, mostly on individuals (but also households, companies etc.) (Quandt and Mauer 2012). These data are chiefly, though not exclusively, gathered by large scale survey programmes and through official statistics (Quandt and Mauer 2012). They have also been described as secondary data, because apart from them being analysed by data collectors or primary researchers, they are suited to be analysed in secondary research. A growing share of these data are not even analysed by data collectors because they are gathered explicitly for broad use by as many researchers and in as many contexts as possible (“secondary” primary data) (Lewis-Beck 2004b, 1009), for example the General Social Survey or the British Social Attitudes Survey (Clark and Maynard 1998).

Because of their impact, the use of these data – before and henceforth simply called survey data – is in focus of the present investigation.

2.1.1 The Specifics of Survey Data and Data Archives

Technically, survey data are numeric data that can be analysed using statistical software (Corti 2004). They are captured data files that consist of matrices which contain values of variables (arranged in columns) for each case or unit of analysis (arranged in rows) (Bernard 2013), also referred to as "case by variable grid" (Bulmer, Sturgis, and Allum 2009, XXIV). The values are numbers that represent manifestations of variables for each case. Which manifestation is represented in which number (or code) is to be defined in additional material, for example in a codebook. Without this additional information, a survey dataset is not interpretable by humans or machines:

> "A data file is ultimately just a string of numbers and not understandable on its own; it can only be interpreted and comprehended intellectually through use of the technical documentation, which indicates a variable's location in the numeric data file, the question it was based on, all possible responses to the question, how the population of interest was sampled (for surveys), and so forth. Together, the data file and its documentation make up the Content Information, sometimes called a data collection or a study." (Vardigan and Whiteman 2007, 76)

This is why *documentation* in the form of metadata is of critical importance in data reuse. Codebooks and other technical documentation can be categorized as variable-level metadata, alongside study-level metadata (describing the context of the study and details of data gathering), file-level metadata (describing data file properties), and administrative and structural metadata (mainly important for maintenance and preservation) (cf. Gutmann et al. 2004).

Users of survey data need documentation that allows them to estimate trustworthiness, data quality and integrity, and relevance of the data for their own purposes (Cf. Carlson and Anderson 2007; Faniel, Kriesberg, and Yakel 2016). Over the last decades, social science data archives have been set up in many countries to meet these needs of secondary survey data users. These institutions play a major supportive role in data seeking and data use practices of survey researchers. Their main goals are to foster data sharing, to preserve data for future research and reference, and to develop archival standards and promote usage of these standards (Nielsen and Hjørland 2014). Accordingly, the data archives' main tasks include: data collection; data processing; data documentation; providing access to data; providing training in data reuse as well as in general empirical methodology; long term archiving of

data (Gutmann et al. 2004; Scheuch 2003). The data archives are staffed with specialized professionals who are known as data archivists, data librarians, data curators, or the like. These professionals perform the core archival tasks mentioned above and are reference persons for people who are looking for and using archived data. They offer different ways of communication, for example direct ways like e-mailing or indirect ways like dissemination of online information or leaflets etc. and, of course, data documentation. Because of their important role as intermediaries in data seeking, data archive staff are interviewed in the qualitative part of this study.

One of the most important task of data archive staff is documentation. Data archives invest considerable resources in documentation (Gutmann et al. 2004). In the last decades, standards and best practices in social science data documentation have been developed by the community, most prominently in the form of the Data Documentation Initiative (DDI)[8]. The DDI presents an international metadata standard that is in use by a majority of social science data archives (Gutmann et al. 2004). In its latest version, the DDI offers documentation along the data lifecycle, that is to say, it enables standardized description of study concept, data collection, data processing, data archiving, data distribution, data discovery, data analysis, repurposing, and data archiving (Vardigan, Heus, and Thomas 2008). Obviously, the DDI is a powerful XML standard that allows for very detailed documentation. However, the full potential of this standard is rarely exploited. Apart from large survey programmes, data from social science studies tend to be superficially documented. Therefore, as Niu and Hedstrom (2008) have shown, social scientists depend on additional information that they retrieve from colleagues, from literature, and other sources; they consider these information activities as a natural part of their work as researchers (Kern and Mathiak 2015).

Through their work, data archive staff have a rich knowledge of contexts, problems and barriers of data reuse. This is why the present study seeks to infer a theory of data seeking behaviour from interviews with these specialists. As intermediaries, data archive staff possess relevant tacit knowledge to produce a grounded theory.

[8] https://ddialliance.org/, accessed October 5, 2020.

2.1.2 Data Seeking During the Survey Research Process

Just like discovery and use of any other information source, discovery and reuse of survey data are influenced by the whole social science research process (cf. Bouazza 1989). Survey research is a key methodology in various social sciences such as sociology, psychology, and political science. These disciplines investigate human behaviour (and thoughts) at individual and group level (Bernard 2013). In line with the social constructivist perspective held in the present study, this research process is not understood as a particular procedure employed by an individual, but in the sense of practices adopted in the research community (Cf. Recker and Müller 2015). Scientific communication in these fields is shaped by prevailing historical and cultural norms. For instance, the affiliation to a particular school of thought may lead to restricted consideration of relevant research (Borgman 2007).

In the literature there are various more or less detailed accounts of what constitutes empirical social research. Some of them are given in the form of flow charts, for example by H. Russel Bernard (2013) who depicts the ideal research process in four consecutive stages: (1) Problem; (2) Method; (3) Data Collection and Analysis; and (4) Support or Reject Hypothesis or Theory (Bernard 2013, 62). There are several problems associated with accounts of this type. For a start, as Bernard points out himself, research does usually not follow this neat sequence, but is often messy (Bernard 2013). What is more, social research is particularly multi-variant in its strategies and approaches, which is why individual activities do not necessarily occur in every research project or at least not in the same sequence (Bryman 2012). However, for analytic purposes it remains helpful to identify discrete research activities. It is not necessary though to stick to a sequential model. An alternative approach is given by Alan Bryman (2012) who prefers to introduce key elements that occur in most social studies: (1) Literature review; (2) Concepts and theories; (3) Research questions; (4) Sampling cases; (5) Data collection; (6) Data analysis; and (7) Writing up (Bryman 2012, 8–15). For the present context it is assumed that all these elements potentially influence data seeking behaviour and specific assumptions on the influence of each of these elements are made in the following paragraphs. It is worth noting that the secondary research process is specific in that it refers to “a set of research endeavours that use existing materials” (Kiecolt and Nathan 1985, 10).

In case of the (1) Literature review, it is possible that evaluation of literature on a topic of interest directly leads to a specific dataset that has been analysed and cited by a given author (similar to the information seeking activity of *forward chaining*, identified by Ellis). This can happen either purposefully or serendipitously. A special case of *chaining* from literature to data is related to the relatively new development of so-called data journals that are introduced especially to describe datasets that are available for reuse.[9] A lead to the applicability of the characteristic of chaining to data seeking can be found in in a recent study by Dagmar Kern and Brigitte Mathiak. The authors report that the studied users of a data catalogue pointed out the importance of data-related literature (Kern and Mathiak 2015). In a study of ecologists who are looking for data, Ann Zimmerman (2007) found that researchers' general familiarity with literature and research trends has a positive influence on their data seeking processes. Additionally, if a researcher has already found interesting data, they can consult related literature in order to find out about general potential, suitability, possible problems, barriers or other issues concerning the dataset in question (given that there are means of identifying related publications, that is to say mechanisms for *backward chaining*). The catalogues of social science data archives usually include related literature on their holdings. For example, the data catalogue of the Inter-university Consortium for Political and Social Research (ICPSR) contains a record for the General Social Survey, 1972-2010 [Cumulative File] [10], which includes about 90 related publications. Furthermore, Ixchel Faniel and colleagues (2013) found out that quantitative social scientists use bibliographies to find literature that helps them make research decisions or gives them more details on measurements or methods (Faniel et al. 2013). Review articles concerning the datasets or articles on previous reuse were also accounted for by secondary researchers (Faniel et al. 2013). Finally, if a researcher is working permanently with a specific dataset, because it is a key resource for their field of interest and offers substance for an infinite number of research questions, they might want to keep track of other research that has been done using this dataset (and hence they perform the information seeking activity of

[9] The first data journal for social research has been created by the Dutch data archive DANS (Data Archiving and Networked Services). It is called Research Data Journal for the Humanities and Social Sciences and publishes articles that aim at "putting the data in a research context" (https://dansdatajournal.nl/rdp/index.html, accessed October 5, 2020).

[10] http://doi.org/10.3886/ICPSR31521.v1, accessed October 5, 2020.

monitoring, for example by checking the updates of the publications list of the dataset record in a data catalogue).

(2) Concepts and theories can be both prerequisites and results of research (Bryman 2012). They are necessary to organise and communicate research design and findings (Bryman 2012). With regard to concepts and theories, data should either be suitable to "shed light on a concept" (Bryman 2012, 9) or yield new concepts that contribute to answering a research question (Bryman 2012). When concepts and theories are guiding research (deductively), they can be seen as an important *context* factor in information seeking behaviour. There is a strong relation to the school of thought a researcher belongs to and to the methodological approaches that they favour. For data seeking this means that, presumably, individuals and institutions holding relevant data are likely to be found within the community of researchers that subscribe to the same theoretic direction (again, there is a lead to the personal *invisible college*). With regard to concepts and theories that might emerge from research (inductively), researchers need data that are interpretable in this direction. This means that they should ideally have been collected independently from strong theoretic or conceptual boundaries. Detailed *documentation* plays an important role in these cases, because it informs secondary researchers about relevant circumstances of the data collection relating to design, fieldwork, and methods.

The (3) Research question can be seen as a general *task* or *problem* that data seekers are facing. Research questions form a particular close relationship with research data. For example, when collecting their own data, research questions will guide the development of the measurement (questionnaire), the sampling, and the analysis of the data (cf. Bryman 2012). Similarly, with regard to data seeking for secondary analysis, the interplay of data and research question is ever-present. Data seeking with a particular research question in mind is always guided by the relevance criterion of the data being suitable to yield answers to the question. Given that secondary analysis is "the reanalysis of existing survey responses with research questions that differ from those of the original research" (Nathan 2004, 1008), it is most likely that during the data seeking process, the research question is more or less revised, according to what is available. Another possible approach is that, coming from an interest in a certain topic rather than from a specific research question, a secondary researcher finds interesting data on this topic which inspires their development of a research

question. On a more general level it can be said that for secondary researchers it is advisable to “merge general substantive interests with familiarity of existing data” (Nathan 2004, 1009). That way, divergence of research questions and available data will not be a major issue. Especially for the more experienced social scientist this merger will probably be natural characteristic of their research work. In her study on how ecologists are proceeding when they are locating data for reuse, Ann Zimmerman (2007) presents patterns of use that suggest an in-between approach. She describes the process of data seeking as starting out with a research problem, formulating research questions, and developing search criteria from there. According to Zimmerman, researchers should be equipped with a sense that data exist and the ability to pose research questions with relation to these data, before they actually start searching for reusable data. She comes to the conclusion that experience with data collection as well as familiarity with literature and research trends positively influence these abilities.

(4) Sampling cases and (5) data collection are research activities that are not conducted by secondary researchers, since they have been done by others, usually the primary researchers. This situation results in several problems that the secondary data user is confronted with: lack of familiarity with the data; complexity of the data; no control over data quality; and absence of key variables (Bryman 2012). To meet these problems, secondary researchers have to invest considerable time and work in evaluation of available data. David W. Stewart and Michael A. Kamins (1993) suggest that they proceed by seeking answers to six questions: What was the purpose of the study? Who collected the information? What information was actually collected? When was the information collected? How was the information obtained? How consistent is the information with other sources? (Stewart and Kamins 1993) In any way, the secondary researcher will have to determine certain criteria concerning sampling and data collection that the survey data they will use for their analyses should meet. Because this data is already existent, the secondary researcher has no influence on the drawn sample, the design of the measurement (questionnaire), or the coding frame. Hence, instead of sampling cases and collecting data, the secondary analyst of survey data will have to run several checks to ensure that the sample represents the population of interest, that the data include the variables needed (validity of the data), and that the unit of analysis is appropriate (Nathan 2004). With regard

to data seeking this means that, apart from topical matching, these sampling and measurement criteria should be critical *relevance criteria* in data retrieval. In her study of ecologists' data seeking behaviour, Zimmerman (2007) identified various relevance criteria, including a match with field-specific standards of data gathering (e.g. with regard to representative sampling) as well as time and geographic information. Zimmerman found that, apart from topical relevance, accordance with scientific practice, data accessibility, and data quality were important criteria for data reuse in ecology. Similarly, Ixchel Faniel and colleagues (2016) found in a study of data reusers in the social sciences that data quality attributes significantly contributed to the scientists' satisfaction with reuse. Overall, since exact matches may be difficult to achieve, secondary researchers should always "be creative in their approach to combining data both within and across surveys, and by incorporating data from outside the data set" (Nathan 2004, 1009). Of course, this requires respective skills and knowledge.

(6) Data analysis can be viewed as the core activity in survey research. It refers to various procedures of managing, statistically analysing, and interpreting the data at hand (cf. Bryman 2012). Ideally, all data seeking activities should have been performed prior to this phase of research. However, there is the possibility that data analysis yields a demand for different or additional data. Especially when working with very complex, hierarchical data, a need of different or additional variables can arise in the course of the analysis. In that case, the secondary researcher might want to start another data seeking process. A specific information seeking activity that can occur during data analysis is the *verifying* characteristic (as identified by Ellis in chemists' information seeking behaviour, see subchapter B.1.2). Verifying might occur if, for example, the data analyst decides that an identified relationship between two variables needs confirmation by one or more additional variables (multivariate analysis) (cf. Bryman 2012). What is more, even though many datasets that are especially suited for secondary analysis are of high quality, there can be no guarantee that they are free of errors. For these cases, good documentation includes a list of *errata* that informs about known errors in a dataset and points to revised versions.

The (7) Writing up usually concludes a research project, and typically several information seeking activities form part of this phase. With regard to data seeking, the main activity in

this stage should again be *verifying*, in particular to check whether there has been a new version of the analysed dataset.

2.2 Needs, Purposes, and Barriers in Survey Data Seeking

Information seeking has been defined as purposeful with regard to a specific information need (Savolainen 2017; Wilson 2000). Needs and purposes are important concepts to include in theoretical assumptions on data seeking. Needs, purposes, and even barriers arise from contexts of data reuse. In the case of the researcher looking for survey data, the context in which needs and purposes arise is made up largely by the domain of empirical social research and the research process described above. The described characteristics of survey data are other influencing factors, as are research trends, such as the prevalent interest in studies of change, that are hardly conductible without using precollected data (Kiecolt and Nathan 1985). Additionally, methodological trends determine further criteria in data seeking. For example, in the case of survey research, the more advanced researcher will lean towards comparative analyses and hence need more complex data sets (Scheuch 2003). The discussion of Bryman's key elements of social inquiry (see above) presages that methodology plays a vital part in survey research. And indeed, a peculiarity of social research that shapes communication in the field is the education in and use of methodology. It is distinctive of the social sciences that instruction in methodology (research design, data analysis) forms a respectable part of the curricula (Borgman 2007). On top of that, research in methodology constitutes a field of its own within the social sciences, populating specific journals and conferences.

On these grounds, the *need* to use existing survey data arises from different levels corresponding to the kind of problem that a researcher is about to solve. These information needs can be categorised in (1) a need to answer new empirical research questions; (2) a need to advance theory; (3) a need to advance methodology (cf. Sieber 1991). In addition, secondary data can serve (4) the need to "verify, refute, or refine original results" (Sieber 1991, 11); and (5) the need to illustrate findings or methodology in teaching (Clark and Maynard 1998).

With regard to these needs imaginable *purposes* that can be served by survey data are, for example:[11]

- Consult classic studies to identify research problems and develop new research questions
- Extract new subsamples or cross-sections from data sets to answer new research questions
- Combine data from independent surveys to answer new research questions
- Combine new data with old data to answer new research questions
- Combine data sets to analyse intercountry differences or to study change over time
- Combine data sets to compare findings across time or locations
- Combine data sets due to small samples in individual studies (by employing internal replication or pooling)
- Combine existing survey data with data from another level of aggregation (multi-level analysis), for example with data from official statistics
- Analyse results from diverse collections to develop new research questions
- Develop a new theory from existing data
- Test another theory on existing data
- Try new hypotheses on existing data (including preliminary tests of hypotheses in an early stage of research)
- Try other methods on previous research using the same data
- Try other models on previous research using the same data
- Generalize, extend or revise findings from previous analyses of the data, for example by including previously unused items
- Repeat analyses from a previous study from within a different theoretic framework
- Analyse data in an exploratory way to inform the design of new data collection, for example with regard to sampling, questionnaire design, or issues of operationalisation
- Analyse existing data in an early stage of research to assess efficacy of and refine or improve measures or questions

[11] This collection of scenarios is informed by various contributions on secondary analysis (Bulmer, Sturgis, and Allum 2009; Clark and Maynard 1998; Clubb et al. 1985; Gould and Handler 1989; Kiecolt and Nathan 1985; Law 2005; Medjedović 2014; Sieber 1991).

- Analyse existing data to inform the conduct of a follow-up study
- Repeat calculations from a previous study to verify the results, for example in a journal review process
- Instruction, training, teaching (for example in statistics)
- Contribute to general knowledge on processes and structures of social or psychological change
- Contribute to methodology, for example by delivering harmonization of data and items across countries or by testing new methods on existing data

An important aspect is the occurrence of *barriers* in data seeking, as they are conveyed by the requests. Apart from the obvious problem of missing data on a given subject, population etc., possible issues in data seeking may be found in the realm of legal questions, for example of personality rights or copyright. Difficulties with data quality, technical data access, documentation, or data complexity may pose barriers, too. Survey data are more complex than data in other disciplines (Curty et al. 2016), which makes data literacy an important factor in data reuse and probably also in data seeking. As Ann Zimmerman (2007) in her study of ecologists found out, experience in data collection has positive effects on finding reusable data. Precisely because of the complexity of survey data, dimensions of experience such as data literacy and general experience in survey research should be beneficial in data seeking. Zimmerman (2007) also found that the studied researchers perceived inadequate documentation as a barrier in their data seeking as well as the general fact that not all primary researchers were willing to share their data for reuse. Poor quality of data and of documentation had already been named as relevant barriers by Kathleen Heim (1980), who did a long term study of users of a social science data archive. She also found evidence for the need of a "centralized inventory of data" (Heim 1980, 225). Until today, such an inventory does not exist. Given the widely distributed research infrastructures and permanent increase of data production, this idea seems overambitious. A 2016 qualitative study of social scientists by Renata G. Curty found that data reusers acknowledge the effort that is required to find suitable data (Curty 2016). It seems to be a more realistic estimate of research reality that secondary data users have to accept and learn about the various ways of data discovery. A less ambitious demand than developing a "centralized inventory" would be to provide better indexing of survey data which would

improve its findability. Subject indexing for survey data is neither applied extensively nor standardized in any way (Friedrich and Kempf 2014). For systems design it would be helpful to know, whether this circumstance is perceived as a barrier in survey data seeking by secondary data users.

With regard to comparative studies or analyses of change, problems with item and sample comparability are to be expected (Kiecolt and Nathan 1985). Furthermore, when trying to operationalize concepts that have not been subject to previous analyses of the existing data, finding the right combination of items can be difficult (Kiecolt and Nathan 1985). Presumably, factors like insufficient training or computer infrastructure (Tenopir, King, Spencer, et al. 2009) will be of less relevance, but they might still occur. An objective of the qualitative study with regard to barriers was to find out from the experts' perspective whether there are barriers that re-occur.

2.3 Specific Factors Influencing Data Seeking

Until today, studies in information behaviour, particularly with regard to scholarly environments, have largely focused on literature use – research on data seeking behaviour still is relatively sparse. Thankfully, in recent years, there have been an increasing number of studies analysing data use. These studies are often conducted in the light of the hot topic of *data sharing* which includes both sides of the same coin: providing data for reuse on the one hand and reusing provided data on the other. Most investigations concentrate on the former phenomenon, enquiring, for example, primary researchers' (lacking) motivation for providing access to data (e.g. Kim and Stanton 2014; Tenopir et al. 2011). Determinants of data use in the person of the secondary researcher have been less in focus. Only recently, research in data reuse is somewhat increasing, probably because the concept has gained importance with regard to funding policies (van de Sandt et al. 2019). Oftentimes studies that investigate the usage side of data sharing are concerned with questions of motivations or intentions for data reuse, quality assessment, and making sense of data (e.g. Curty 2015; Curty and Qin 2014; Faniel, Kriesberg, and Yakel 2016). Notable exceptions are the works of Ann Zimmerman (2007), Ixchel Faniel et al. (2012, 2013), Ayoung Yoon and Youngseek Kim (2017), and Ayoung Yoon (2017). Findings from these and other information behaviour studies that did consider research data use and point to specificities with regard to data seeking will be presented in the following paragraphs.

2.3.1 The Importance of Data Documentation

Before the 1990s, survey data or other research data had been acknowledged in a few investigations of social scientists' information behaviour (e.g., in the INFROSS study in 1967 and in a meta-analysis by Michael Brittain in 1982). These studies were very general in their results regarding survey data reuse. In 1991, Stephen Stoan likewise concluded from a literature-review of academic information retrieval, that data as an information source was of general relevance for the social sciences. The specific problem that Stoan identified was that these materials lacked indexing, were difficult to obtain and to assess in their relevance (Stoan 1991). This assessment points to a very important specificity of research data seeking, which is the importance of *documentation*. As has been explained above, research data, in particular survey data, are oftentimes not interpretable without documentation. What is more, even if documentation is provided, the detail and quality of documentation is what really matters with regard to reusability of data. In 1980, Kathleen Heim delivered the first in depth study of information seeking and use of social science data users. The site of this study was the Data and Program Library Service (DPLS) at the University of Wisconsin-Madison ("a highly developed archive of the general-purpose, local service type" (Heim 1980, 211)), and users were studied over a period of ten years. Heim showed that poor quality of documentation was indeed a relevant barrier in information seeking and reuse for her respondents, as was poor data quality and the lack of a "centralized inventory of data" (Heim 1980, 225).

More recent research confirms the continuing relevance of data documentation in data seeking and reuse. With regard to novice social scientists, Ixchel Faniel and colleagues (2012) found that documentation that provides information on the primary researchers' research process was of major interest to the secondary users, for example "fine-grained details about the data collection and coding procedures" (Faniel, Kriesberg, and Yakel 2012, 7) such as questionnaires, their development and employment. In a follow-up comparative study of data reuse among social scientists and archaeologists, (Faniel et al. 2013) the authors elaborated somewhat further on the information on the contexts of data production. In general, the studied quantitative social scientists tended to reuse survey data, be it from individual or institutional principal investigators (Faniel et al. 2013). Context information on the data was found to be provided in digital and static form (for example, accompanying text

documents or survey material) (Faniel et al. 2013). Information that the secondary data users were particularly interested in were “the data producer’s research methods, especially […] insight into how the data producer carried out the research” (Faniel et al. 2013, 798). For example, the users “wanted to know how data producers defined and measured the variables data were intended to capture” (Faniel et al. 2013, 798). In a study of users of data provided by the Inter-University Consortium for Political and Social Research (ICPSR), Ixchel Faniel et al. (2016) confirmed that the social scientists' satisfaction with data reuse was positively influenced by documentation quality.

The key role of documentation for survey datasets has also been in focus of a recent retrieval study by Dagmar Kern and Brigitte Mathiak (2015). The authors designed their investigation around the same data retrieval system that was used for sampling in the present study, the DBK (Datenbestandskatalog) hosted by the data archive at the GESIS Leibniz Institute for the Social Sciences in Germany. Kern and Mathiak conducted a lab study with “simulated work task situations” (Kern and Mathiak 2015, 198) employing the DBK, and telephone interviews with DBK users. The study had 53 participants, including professors, undergraduate and graduate students, as well as postdoc researchers, mainly from the fields of social sciences and political sciences. The main result of this study was that – different from the situation in literature retrieval – DBK users did not mind to invest much time in the study of the documentation in order to make their relevance judgement. Apparently, they deemed this investment to be a necessary part of the research process. Kern and Mathiak stress the importance of documentation for data retrieval and point out how additional material (codebooks, method reports, survey instruments) contains essential information without which data are useless (Kern and Mathiak 2015). The authors conclude that researchers proceed differently when they are searching data than when they are searching literature in that they invest considerably more time in documentation scanning and relevance evaluation. Kern and Mathiak come to the assumption that this is because “data sets are much more decisive for research activity than any literature is” (Kern and Mathiak 2015, 207), and “time spent on choosing the correct data set is therefore time well-spent” (Kern and Mathiak 2015, 207). Another obvious reason, however, should lie in the complexity and variety of research data as an information unit as compared to literature.

These findings stress the importance of documentation in survey research and help to explain why social scientists tend to invest much time and effort in documentation scanning. In contrast to these results, Jinfang Niu and Margaret Hedstrom report in their 2008 and 2009 studies how data archive staff experience that "typical questions that users raise sometimes are already answered within the documentation accompanying the data" (J. Niu and Hedstrom 2008, 3). Niu and Hedstrom's research on the role that data documentation plays in information seeking behaviour of survey data users sheds light on how users actually perceive and use documentation. The authors have designed a Documentation Evaluation Model for Social Science Data (DEM) (J. Niu and Hedstrom 2008) and tested this model using data that they collected by asking social science researchers to judge data documentation (J. Niu and Hedstrom 2009). The study is based on a quantitative survey and in-depth interviews with researchers who conduct secondary data analysis in social science research. Main results of the investigation include (J. Niu and Hedstrom 2009, 128):

- Poor documentation is regarded as poor no matter how experienced the users are
- Good documentation is more sufficient for experienced users than novice users
- Documentation of data produced for sharing is good in general but not perfect
- Perceived documentation quality varies with the characteristics of data
- Perceived documentation quality is weakly affected by users' absorptive capacity

In conclusion, characteristics of data had a clear impact on perceived documentation quality, while factors that lay in the person of the user only weakly affected how they judged the documentation (J. Niu and Hedstrom 2009). From the perspective of information behaviour research the most important finding from this study may be that documentation quality is not as relevant for the decision to use a certain dataset; in fact, researchers deem it necessary to gather information beyond documentation anyway in order to use the data properly, especially when they are less experienced and/or using data that has been produced specifically for secondary use (J. Niu and Hedstrom 2009). Niu and Hedstrom's study suggests that poor documentation has a higher impact on the decision whether to use datasets produced by other individuals than on the decision to use data produced by large survey programs. This shows that, even in the digital age, researchers' information seeking behaviour is influenced by more than just documentation and information retrieval systems, even if documentation can be said to be quite good. Sure enough, documentation is critical,

but secondary researchers interested in certain datasets use different methods of acquiring information on the data, such as reading related literature, browsing websites of data producers, or asking experts and intermediaries such as data archive staff or even other secondary data users (J. Niu and Hedstrom 2008). The authors indicate that users even tend to consult data archive staff although answers to their questions could have been found in the documentation (J. Niu and Hedstrom 2008).

2.3.2 The Role of Intermediaries

This result points to another specific characteristic of data-related information seeking, which is the role of intermediaries in the process of seeking and reusing data. In her 1980 study, Kathleen Heim already found out that the surveyed data archive users were most likely to find out about needed data from colleagues. Another study illustrating the importance of intermediaries and informal ways of data seeking is the one conducted by Carol Hert and Gary Marchionini (1997, 1998). Their work focused on seeking statistical information and involved usage of three different federal websites. The most valuable finding of Hert and Marchionini's research refers to the manifold and important roles that intermediaries play in data seeking. It is apparent from this study that, as long as finding and using complex data requires a large variety of additional information, coaching and consultation, data users will continue to rely on personal contacts. In the expert interviews conducted by Hert and Marchionini, intermediaries named a whole range of activities that they performed to aid user requests: pointing users to sources; understand user queries through reference interviews, use of domain knowledge, previous queries; explaining data collection strategies; helping in interpretation of statistics; providing technical assistance (downloading data, data import, computer based searching); remaining abreast of the existence of data sources, locally, regionally, and nationally; networking with other agencies or people to which they can refer users (Hert and Marchionini 1997). These insights are very interesting for the present study that also includes expert interviews with intermediaries.

Faniel and colleagues (2012) found that novice social scientists in particular rely on more experienced researchers when looking for and reusing data. The more recent study by Faniel and colleagues (2013) revealed that data reusers in the social sciences in general relied on the help of peers as intermediaries with regard to finding data and judging their appropriateness (Faniel et al. 2013). These results show that the importance of informal

information channels as it has been identified repeatedly in past studies on social scientists' information behaviour (e.g., the APA studies, INFROSS, etc.) is verifiable for the special case of data reuse.

With regard to the relevance of intermediaries in data reuse, the results from Faniel and colleagues are in line with the more recent findings from Ayoung Yoon and Youngseek Kim (2017) and Ayoung Yoon (2017).[12] Yoon and Kim (2017) studied survey data reuse behaviours in the social sciences. They surveyed 292 researchers that they had sampled from an online science database. The authors found that perceived effort in data reuse was a relevant factor in the scientists' motivation to reuse data (Yoon and Kim 2017). As Yoon and Kim go on to explain, social science data archives (or data repositories, as they call these institutions) play a critical role in reducing the researchers' perceived effort to reuse data. The data archives are credited with providing access, documentation, error management and added value, which help data users to find and reuse data. Even though these institutions cannot provide solutions for all problems (such as unpublished data or imperfect documentation), Yoon and Kim stress their continuing importance for the data reusing community. To these findings Ayoung Yoon adds even more relevant results with her 2017 investigation of quantitative data reusers from the fields of public health and social work (Yoon 2017). In this qualitative study with 38 interviewees, Yoon investigated the role of communication in data reuse practice. She found out that in the process of reusing data, the studied researchers communicated with various people to receive support in the areas of searching data, learning about data, and solving data reuse problems. The participants in this study reported that they would reach out to peers, supervisors or data professionals at various stages of their research process. The main result of the study is that communication with different communities supports searching, learning and problem solving processes in data reuse. These findings clearly point to the importance of contacting intermediaries when looking for data. The intermediaries that are important in this context are not only data archive professionals but also the people who have collected the datasets of interest (principal investigators).

[12] Both studies were published after the qualitative data collection of the present study was finished.

2.3.3 The Role of Information Technology and Automation

Another result of the already mentioned study by Hert and Marchionini (1997) was a typology of users of the three analysed websites. The resulting user profile was as diverse as it gets as it included business users, academic users, the media, general public, government users, education users, statisticians, users at libraries, museums and other non-profit users (Hert and Marchionini 1997). Hert and Marchionini suggest that "[…] designers [could] use the taxonomy as a framework for implementing alternative user interfaces" (Hert and Marchionini 1998, 307) and that "[t]he taxonomy can mediate user-system communication to facilitate access to the data appropriate to a specific task(s), or have the system suggest possible strategies for exploring and using the site" (Hert and Marchionini 1998, 307). From today's point of view this analysis seems rather theoretic and feasibility as well as usability of respective web services is questionable. The ideas seem to overestimate the share of problem solving to be contributed by the system. Sure enough, users of web services benefit from technical support, but only as long as they don't feel too preoccupied with handling the system. However, it remains an interesting question, to what extent current problems in data seeking can be solved or otherwise addressed by IT solutions. Is it possible that, with the advent of digital information provision, the relevance of informal information channels and intermediaries in data seeking has decreased (Vakkari and Talja 2006; cf. K. Fisher and Julien 2009)? Hert and Marchionini concluded that: "The availability of such systems has meant that the potential for people to find information without the help of knowledgeable intermediaries has increased." (Hert and Marchionini 1997 n.pag.)

At least with regard to seeking literature it can be stated that the increase of electronic versus print journals in all academic fields had a major impact on information seeking (cf. Athukorala et al. 2014; Tenopir, King, Edwards, et al. 2009). Carol Tenopir and colleagues in their longitudinal study covering 30 years of scholarly activity conclude:

> "With the growth of electronic journals, the continued increase in the number of journals and articles published yearly, and alternative sources of scholarly articles, many information seeking and reading patterns of science faculty are changing. Articles are identified and located through a variety of information-seeking methods, including browsing, online searching, following citation links, and getting recommendations from colleagues, yet the proportion of articles located by searching is increasing." (Tenopir et al. 2010, 26)

With regard to scientific data, Bradley M. Hemminger and colleagues (2007) who studied information seeking behaviour of science faculty found that finding information in the digital age had become so easy that online searching was employed routinely and successfully regardless of the type of information material, be it a journal article or a genetic sequence (Hemminger et al. 2007). They conclude that "this type of access has surpassed personal communications, and it is close to journal articles in frequency of use by researchers" (Hemminger et al. 2007, 2214–15). One reason why this may not be the case for social science data access is that the latter is more complex in terms of structure and exists in more types and variations than the data collected in a particular scientific discipline (Robbin 1995). As Blaise Cronin in his seminal paper on invisible colleges explained:

> "For [social scientists] information has a variety of meanings and forms (e.g. published research results, experimental data, time series, field work findings, data files, archival data, precedents, patent information, original manuscripts, oral history) and it seems reasonable to assume that the kind of information which is required, the ease with which it can be accessed and the use to which it is likely to be put will have a direct bearing on the way in which interpersonal networks are developed and relied upon." (Cronin 1982, 230)

Assuming that the role of intermediaries in data seeking prevails, does not mean that information technology has had less impact on this information practice than on others. For instance, with regard to information seeking in the social sciences, Xuemei Ge's (2010) empirical study confirms that digital environments influence seeking, access and use of information sources (Ge 2010). Moreover, digital availability of sources seems to increase the demand for more electronic information (Ge 2010). Also, the influence of the social web is apparent in today's information practices of social scientists (Ge 2010).

It seems plausible that developments in online communication rather furthered possibilities of exchange between researchers and intermediaries in that they "extend the opportunities for interpersonal information seeking." (Borgman 2007, 157; cf. Ge 2010) Colleagues and other people still serve as direct information sources (e.g. by answering questions) or as pointers to information, for example to data (Borgman 2007). Only the means of communication between researchers have increased through availability of e-mail, online social networking etc. (Borgman 2007). With regard to a possible replacement of personal networks (*invisible colleges*) through information technology, Cronin hypothesized more

than 30 years ago that "[t]here seems little doubt that developments in communications technology will herald a new mode of invisible college, but it seems unlikely that electronic networks and tele/video conferencing will in the short term entirely replace conventional communication channels." (Cronin 1982, 232) Current analyses of the influence of information technology on researchers' general information behaviour have been plenty and conclusions range from those judgements that see historic impact (Krampen, Fell, and Schui 2011) to those holding that the influence should not be overrated (Bates 2010). In any case, information technology influences the "information environment" of researchers, and hence is a relevant contextual factor of scholarly information practices (Ge 2010).

3. Areas of Exploration

The present study aims at building a grounded theory of data seeking behaviour based on qualitative interviews and exemplifying this theory by a quantitative survey. In preparation of this study, the purpose of this chapter was to generate theoretical assumptions that submit to the existing research that has been discussed above. These theoretical assumptions don't comprise a coherent theory of data seeking behaviour, which is why they can only serve as a background to the grounded theory that remains to be developed. The following paragraphs summarize the most important theoretical assumptions that inform the areas of exploration for the qualitative study.

As we have seen in the previous sections, there are good reasons to assume that research data seeking behaviour differs from literature seeking behaviour and is domain specific (in the sense of Hjørland and Albrechtsen). Consistent with the social constructivist viewpoint, the key concept of *context* provides the frame of reference from where individual theoretical assumptions are made. Understood as "the place where meaning is socially constructed" (Tabak 2014, 2225) *context* is not made up by a given set of factors ("context as a container" (Courtright 2007, 286)) but "is embedded in action and practices" (Savolainen 2009, 39). Context factors are to be found in research processes and standards as they prevail in the domain of survey research. These factors are apt to render relevance requirements regarding data quality, topics and methodology when looking for data.

The present study also focuses on individual factors in the person of the researcher. First and foremost, the users' needs, goals, purposes, and requirements are to be considered. Special interest is given to the question what factors on the level of the individual are noticeable with regard to specific problems, needs, behaviours, or even barriers. For example, the level of education was expected to be an issue, since it can be assumed that a majority of students and junior researchers refrain from complex data, because they would not need them for their projects (Scheuch 2003). Familiarity with data analysis and experience with data collection are two important aspects of survey data literacy that were expected to have an effect on data seeking behaviour.

Looking more closely at information seeking behaviour specifically with regard to survey data, there are other important factors that need to be taken into account. First of all, documentation seems to play a pivotal role in various practices surrounding survey data seeking and use. Furthermore, the role of intermediaries as well as information technology must be considered as important factors in survey data seeking behaviour.

The theoretical and empirical considerations that have been made in this regard and have been outlined in this chapter lead to seven areas of exploration that will inform the qualitative interviewing. These areas are:

- The users’ educational/professional background
- The users' research experience and data literacy
- Goals, needs and purposes of users
- Requirements (data quality, topics, methods) when looking for data
- The role of documentation
- Information sources and channels (the role of intermediaries and information technology)
- Barriers and problems when looking for data

How these areas were incorporated in the development of the qualitative study is explained further in subchapter "C.2.1 Interview Guide".

C. Qualitative Study

1. Methodology: Model-building with a Grounded Theory Approach

The main aim of the qualitative part of this study was to shed light on the contextual factors of data seeking behaviour. To this end, reference staff from a large European data archive were interviewed, assuming that they are experts on the phenomena in question, having acquired unique and deep knowledge about the (social) contexts of their work field.[13] These reference persons were interviewed in their role as experts in the context of data seeking. Since it has been shown in the past that reference staff are important intermediaries not only in researchers' general information seeking behaviour but all the more in contexts of data re-use (Hert and Marchionini 1997, 1998; Yoon and Kim 2017), it can be assumed that they possess a condensed view of secondary data users' professional contexts, information needs, demands, tasks, goals, and problems. As has been shown above, there is a lack of research on these phenomena, but reference staff in archives possess a rich knowledge in this domain. Raising this already existent knowledge was intended to lead to a better understanding of information seeking behaviour of secondary data users.

It was assumed from the start that, when interviewing intermediaries such as data service professionals, they can only give information on behaviours and practices from those users, who encounter problems when seeking information. This assumption was backed during the field phase when one interviewee stated: „This means, the average user consults us only if they've found an error or believe they've found an error. Or found something missing in the documentation. Or if they are looking for specific data that they don't think they can find." (Interviewee no. 4). This means that all those users who have found sufficient information online or from other sources do not call data service for help. The existing systems, even though they may need improvement here and there, seem to work for them. Given that the results of this study were intended to lead to recommendations for better data infrastructure, focusing on problematic situations that cause data service requests were of vital importance in the data collection. These are the areas, where information systems need

[13] The data archive that served as the setting for the present study is the same institution that the author of the study works at, albeit in different work areas with virtually no cooperation. It is important to note here that this personal involvement of the researcher in the setting cannot be without influence on the conduct and results of the study. The interpretation of the collected data and the resulting grounded theory are rendered by the interpretation of a researcher who is involved in the same setting as the interviewees.

to be improved to make a difference. The investigation of the users' direct perspective was reserved for the quantitative part of the study (see chapter D.).

The qualitative strategy that was employed here aims at collecting data from knowledgeable individuals by reconstructing the (social) situations and processes in which data seeking occurs (cf. Gläser and Laudel 2010). The appropriate strategy is to conduct expert interviews or to *interview informants* (Gläser and Laudel 2010). In this kind of interview, *experts* or *informants* are not the "object of study" – at least not primarily (Gläser and Laudel 2010, 12). They are interviewed as witnesses of the phenomenon under study (in this case: *information seeking behaviour of survey data users*), while their personal life is not in the focus of interest (Gläser and Laudel 2010). This does not mean that the experts' emerging personal thoughts, attitudes, and emotions are irrelevant – they are just not the focal point of investigation. Rather they are important in how they affect the accounts of the reported incidents and impressions. In that respect they are of relevance in data analysis and interpretation (Gläser and Laudel 2010).

Interviewing informants or experts is not associated with one specific technique of data collection or analysis (Gläser and Laudel 2010). A methodology that is compatible with expert interviews and was applied here is the *constructivist grounded theory* approach as it was introduced by Kathy Charmaz (2005, 2014) and is advocated, for instance, by Nick Pidgeon and Karen Henwood (2014). Grounded theory methodology was first introduced by Barney Glaser and Anselm Strauss (1967). In its original form, this qualitative research approach implies the "discovery of theory from data systematically obtained from social research" (Glaser and Strauss 1967, 2). With its proposed techniques of theoretical sampling and constant comparison this methodology has been widely applied since (Bryant and Charmaz 2007) and furthered methodological development in qualitative research in general (cf. Pidgeon and Henwood 2004). The methodology entails "systematic, successive strategies for developing fresh ideas to collect, study and analyse empirical data" (Charmaz 2008, 461) with the objective of creating a (middle-range) theory. Distinctively, the development of grounded theory is conducted by simultaneous data collection and analysis (Charmaz 2008). In particular, (1) specific techniques of coding are applied to the data; (2) coding leads to analytic categories; and (3) refinement and empirical checks of categories lead to theoretical analyses (Charmaz 2008). The whole process of grounded theory development is conducted

by methods of *constant comparison*, which involves several comparative research practices: comparing data with data; labelling data with active, specific codes; selecting focused codes; comparing and sorting data with focused codes; raising telling focused codes to tentative analytic categories; comparing data and codes with analytic categories; constructing theoretical concepts from abstract categories; comparing category with concept; and comparing concept with concept (Charmaz 2008). During this process, further sampling of participants is performed with the objective of collecting more data to refine identified categories and increase precision of the theory (so called *theoretical sampling*) (Charmaz 2008).

Over the decades the implied idea of theory *emerging* from data in the sense of an objective reality has led to criticism and proposals for adjustment (cf. Charmaz 2014). Critics have faulted the "positivist, objectivist direction" (Bryant and Charmaz 2007, 33) or "positivist empiricist philosophy" (Pidgeon and Henwood 2004, 627) behind this idea. *Constructivist grounded theory* is an alternative proposal that has gained broad attention in this context. Kathy Charmaz describes this variant in dissociating herself from Glaser and Strauss' approach:

> Unlike their position, I assume that neither data nor theories are discovered either as given in the data or the analysis. Rather, we are part of the world we study, the data we collect, and the analyses we produce. We construct our grounded theories through our past and present involvements and interactions with people, perspectives and research practices. […] Research participants' implicit meanings, experiential views – and researchers' finished grounded theories – are constructions of reality." (Charmaz 2014, 17)

The distinctive feature of this variant of grounded theory methodology is that it acknowledges the involvement of the researcher as a relevant factor in data collection, analysis, and theory development. Unlike original grounded theory methodology, the constructivist variant acknowledges that researchers do not enter analysis "as a *tabula rasa*" (Charmaz 2014, 306). Consequently – and more realistically – constructivist grounded theory considers the researcher's familiarity with relevant theoretical work (Charmaz 2014). It is important to be aware of any preconceptions one may have due to theoretical perspectives or knowledge acquired prior to data collection (hence the notion of *informed grounded*

theory) and to not let these preconceptions interfere with data analysis (hence the notion of *theoretical agnosticism*) (Charmaz 2014).

A research design aspect in constructivist grounded theory that is particularly influenced by the researchers' background knowledge is initial sampling. This refers to the outset of the study, where the initial participants are selected according to relevance criteria (Charmaz 2014), which are necessarily influenced by background knowledge. Initial sampling is purposive and as such enables the researcher to select from various sampling strategies (Pidgeon and Henwood 2004). The present study started with a sample of key gatekeepers as proposed by Pidgeon and Henwood (2004). These were two reference persons of a large data archive who handle helpdesk enquiries. In their function as primary contact persons in the data service, they help users with general requests and refer them to more specialised colleagues if they have specific questions. These two people were expected to have a broad impression of all kinds of enquirers and enquiries. As the study proceeded, the choice of further participants was guided by theoretical sampling. Different from initial sampling, theoretical sampling should be performed on the grounds of the already collected data; more precisely, the sampling should aim at explication of the categories already identified from the data (Charmaz 2014). How initial and theoretical sampling methods were applied in the present case is explained in depth in subchapter C.2.3.

In practical terms, the constructivist grounded theory approach implies that although researchers bring with them a background of theoretical affiliation and knowledge, at the outset there are no testable theories or hypotheses concerning the research question. Instead there may be key topics or "areas of exploration" (Herring 2013, 206) that can be addressed in data collection, and it is to be expected that guiding aspects might even change due to research developments (Herring 2013). In that regard constructivist grounded theory methodology initially is an inductive approach in that it allows for identification of intriguing cases in the data. As analysis is proceeding, induction may be followed by abduction, in that different possible theoretical explanations for the findings are devised and checked against experience to identify the one that is most plausible (Charmaz 2008). The primary advantage of this way of proceeding is that it allows open thinking and flexibility, which was deemed particularly important for the understudied phenomenon that is dealt with in this study. In the interest of openness and flexibility, the interviews were unstructured and therefore did

not follow a classic interview guide. Instead, they involved questions revolving around the following areas of exploration that have been introduced in subchapter B.3: the users' background; information sources and types; documentation and data quality; the survey research process; research trends; users' needs and purposes; methodological requirements; and barriers and problems. Wherever needed, the participants were asked initiating questions, but in line with the constructivist perspective, these questions were kept as broad and general as possible, "so that the participants can construct the meaning of a situation, a meaning typically forged in discussions or interactions with other persons" (Creswell 2013, 25).

The interviews were audio recorded and transcribed. Based on the transcriptions, open and focused coding was performed (details in subchapter C.2.4), using the methods of constant comparison mentioned above (cf. Charmaz 2014). In constructivist grounded theory methodology the codes are not meant to merely describe the topics of the data, but rather to serve as a means to interpret the data (Herring 2013). They are not just significant keywords that happen to be used by the interviewees but denotations of the participants' actions, attitudes and opinions as they become interpretable during analysis (Herring 2013). Focused codes eventually lead to conceptual categories, for instance by the use of memos (cf. Charmaz 2014). Coding, analysis, and memo-writing induce and determine further theoretical sampling with the aim of validating (or testing) identified categories (Herring 2013). In particular, theoretical sampling is performed to specify and saturate properties of the categories (Charmaz 2014). By analysing and validating the data that way, a number of theoretical codes are produced and developed towards conceptual key categories, which can also be understood as *major categories* (Herring 2013). When saturation of the major categories is reached, these are analysed with regard to their implications and theoretical statements can be constructed accordingly (cf. Herring 2013). Coding, memo-writing, and theory-building of the present study are presented in depth in subchapter C.2.4.

In the present study, the combination of the major categories led to a *grounded theory* of survey data seeking. This theory is grounded in the view of the interviewed experts (Creswell 2014) and stands as a constructed reality of data seeking, as it is interpreted by all participants, including the interviewer. In tradition of information behaviour research, the categories and relationships that make up the constructed grounded theory are described as

a model of data seeking behaviour, including a visual representation (see Figure 14 or, for a larger version, Annex 16). This is a representation in the form of a diagram, which is a frequently used tool in grounded theory methodology (Charmaz 2014; Creswell 2014).

2. Data Collection, Sampling, Coding, and Memo-Writing

2.1 Interview Guide

As noted before, the interviews did not follow a pre-structured interview guide but were designed as unstructured interviews with open questions. Unstructured interviewing has the downside of producing data that makes comparison more difficult. This downside was negligible for the present study that was explorative in character. Comparing of cases needs respective categories (or variables). Since it was the purpose of the interviews to create categories in the first place, comparability of the results was not needed at this point. The main interest was to create rich data to inform the development of a nuanced grounded theory of information seeking behaviour with regard to survey data reuse.

The approach of unstructured interviewing was fruitful in that it turned out to enable the participants to associate more freely and to contribute to the topics from their own point of view and experience. However, even unstructured or open interviewing needs interview topics to ensure that information on the areas of exploration is collected (Gläser and Laudel 2010). The interview guide is needed as a flexible framework that does not determine a fixed order of questions. To this end, an interview guide with open questions was developed. It was prepared in two steps. First, a rough draft was written in English, based on the theoretical assumptions made and the areas of exploration developed in chapter “B. Theoretical Perspective”. These seven areas of exploration were:

- The users’ educational/professional background
- The users' research experience and data literacy
- Goals, needs and purposes of users
- Requirements (data quality, topics, methods) when looking for data
- The role of documentation
- Information sources and channels (the role of intermediaries and information technology)

- Barriers and problems when looking for data

In a second step, the information collected in this draft version was structured and condensed to represent eight interview topics (the areas of exploration that are addressed with each of these interview topics are added in brackets):

- Educational/professional background of data users (The users' educational/professional background)
- Survey data literacy (The users' research experience and data literacy)
- Users seeking support (Information sources and channels; the role of intermediaries and information technology)
- Tasks and goals of users (Goals, needs and purposes of users)
- Relevance criteria and data quality (Requirements when looking for data)
- Research trends (Requirements when looking for data)
- Information and documentation (The role of documentation)
- Problems and barriers (Barriers and problems when looking for data)

This condensed version of the interview guide (Annex 1) was prepared in German, since all interviews would be conducted in German. The interview guide also contains three possible introductory questions that were intended to get the interviewees thinking about their work and their encounters with secondary data users:

- How long have you been working as a reference person?
- How many requests do you get per day or week on average?
- What requests do users have?

The third of these introductory questions is more open than the first two. It was phrased very general in comparison to the eight topics named above, in order to get the participants to set their own priorities.

During the course of the field phase, a closing question was added to give participants the opportunity to add even more to the topic, in case they felt that some aspect had been disregarded. The closing question was: "Did our conversation meet your expectations?" This change was made after the third interview when the interviewee indicated that he would have expected other priorities in the topics addressed. And indeed, this question prompted

interviewees four, five and six to contribute a few more insights. This would remain the only change that was made to the interview guide during the field phase.

2.2 Interviewing

The study includes six main interviews. From the six participants, two were female and four were male. The shortest interview lasted about 52 minutes, the longest 1 hour and 36 minutes. Beforehand, one pilot interview was conducted to test the interview guide, the informed consent form and the circumstances of data collection (see subchapter C.2.2.1). The main interviews were conducted from June 9th to July 6th 2016. By means of respondent validation (see subchapter C.3.3) the developed theory was presented to participants of the study in 2018 in order to validate for accuracy, but also for topicality.

All interviewees were invited several days or a few weeks before the interview took place. Five interviews were conducted in a private meeting room located at the institution where the participants work. One interview was conducted via video call, with the participant answering from their home office. At the beginning of the interview session, all participants were presented with a two-page document (Annex 2) that explained the rationale of the research project and the planned handling of the interview data on page one. Page two of the document was a form that included a declaration of informed consent, to be signed by the interviewee.

The interviewees were informed by the said document and verbally, that two recording devices would be used to record the whole interview. All participants agreed with the interviews being recorded. All recordings turned out to be of very good quality. Copies of the recordings were saved in a secure project folder on a personal network drive. The original recordings on the recording devices were erased afterwards.

Beginning directly after the first main interview, all interviews were transcribed one by one. This work was done with the atlas.ti software, version 7 (ATLAS.ti Scientific Software Development 2012). The audio files were included in the project folder created by the software. The transcript files were included in the same folder. Every audio file was cross referenced with the transcript during the process of transcribing. In the course of transcription, potentially sensitive information such as names of individuals or institutions, project titles, and survey names were flagged with square brackets. After the coding was

finished, the content of these brackets was replaced by general terms such as “person”, “institution”, “project”, “survey”. The interview transcripts only contain these general terms to protect the data privacy of the participants. As stated in the consent form that the interviewees were asked to sign, the audio recordings were deleted after data analysis. The transcripts have been archived and are accessible via the GESIS data archive upon request and for scientific purposes only (Friedrich 2020a).

2.2.1 Pilot Interview

With the objective of testing the adequacy of the planned design of the interviews, especially the interview guide, a pilot interview was conducted. The interview’s content was not analysed further and the included data were not used in the development of the grounded theory. The pilot interview’s sole purpose was to assess feasibility of the study as planned and to identify possible problems with the content or the setting of the interview (Teddlie and Tashakkori 2009).

The chosen interviewee was a former assistant to a data curator. From his experience in this position, he was able to adopt the perspective of someone who is advising survey data users. The pilot interview was conducted on June 2nd, 2016. The interviewing time was 56 minutes and 17 seconds. The interview took place in a private meeting room that was booked for one hour. To prevent disturbances, a sign was placed on the door that said: “Interview, please do not disturb”. Before the beginning of the interview, the participant was presented with the written project information and informed consent form. The interviewee read and signed the document without having any requests. Afterwards, he was informed that the recording device was now being turned on and the interview began.

Since the participant was not presented with a fixed set of questions, he was encouraged to speak and associate freely on the introduced aspects. As it turned out, every area that was included in the interview guide was addressed at some point during the interview. Some of the questions did not need to be asked, because the participant addressed the respective area out of his own narrative. This indicates that the questions or areas listed in the interview guide worked well together. The interviewee was not particularly hesitant in his contributions, nor did he seem to have trouble in understanding the questions. All in all, the interview went very well and as expected. On these grounds, it did not seem necessary to conduct further pilot interviews. Judging from an ex post perspective, the interview guide

had a high validity, because only one question needed to be added during the field phase (the closing question, see subchapter "C.2.1 Interview Guide").

The allocated interviewing time of about one hour turned out to be sufficient as expected. However, when listening to the audio recording afterwards, several instances could be identified where the interviewee was cut off at points where he might have given even more detailed information. In the main interviews, more attention was given to letting participants talk at their speed and giving them some breaks to follow their thoughts. This may be one reason why almost all of the main interviews lasted longer than the pilot interview.

Towards the end of the interviewing time, people were passing the meeting room and looking through the glass door, apparently because they had booked the room afterwards. This was a noticeable disturbance. As a consequence, the room was always booked for one and a half hour for the main interviews. Additionally, the booking time was added to the sign on the door, so people passing by would have an orientation as to how long the room was definitely occupied.

2.2.2 Main Interviews

The main interviews were conducted in the same fashion as the pilot interview. A new copy of the interview guide was printed for every interview. During the interview, notes were made on this copy. The main purpose of these notes was to keep track of the topics that the interviewee addressed during the conversation, because the interviews were not designed to follow a predefined order of questions. The first introductory question was the only one that was presented to all participants at the same point in time in the interviews: directly in the beginning. All participants were asked the second and third introductory questions from the guide as well, but not necessarily in this order. In some of the interviews, participants directly switched to topics covered by more specific questions in the interview guide. It is noteworthy that already the very clear and to-the-point first introductory question prompted some participants to think of more than what was actually asked. It was a high priority to not interrupt participants when they associated freely in that way. As a consequence, not one of the interviews proceeded along the topics or questions as they appeared in the interview guide. When an interviewee addressed one or more topics that were included in the guide, the topic was checked with a pen or noted down with a keyword on the printed copy of the interview guide (an example of an interview guide with notes is

given in Annex 3). Many times participants would talk about specific experiences at length and, in doing so, mention several topics from the list. All these mentions were noted down and taken up again at a later point during the conversation in order to discuss them in depth.

Following this method, all questions in the interview guide were addressed in all the interviews at some point. Not all questions were directly asked in all interviews. Very often, the topics addressed in the questions would just come up in the interviewees' narratives at some point during the interviews.

2.3 Initial and Theoretical Sampling

As explained above (subchapter C.1), theoretical sampling, as a corner stone of grounded theory methodology, was applied in this study. Theoretical sampling is a type of purposeful sampling that proceeds alongside the process of interviewing and data analysis (Pidgeon and Henwood 2004). This sampling strategy entails that during analysis new cases are purposefully chosen with regard to lack of clarity in identified categories in the previously collected data. New cases are drawn until theoretical saturation has been reached (Pidgeon and Henwood 2004). In the present study, this led to a sample of six cases.

The first two cases were drawn by means of initial sampling. In constructivist grounded theory methodology, initial sampling is applied at the very beginning of a study, when no data has been collected or analysed yet (Charmaz 2014). Initial sampling is necessary, because at the outset of a study there are no data-based analytical categories that could be used for theoretical sampling. Different strategies can be pursued to determine the initial sample of cases. The strategy adopted in the present study was "rich response sampling" (Pidgeon and Henwood 2004, 635) of cases that promised a particular broad picture of the phenomenon in question. This sampling was informed by theoretical and practical background knowledge. All interviewees work in the same data archive as the author of this study. This situation provided the study with inside knowledge about the tasks and workflows of the interviewees. With regard to initial sampling, it was useful to know details about the different roles in data service of data centres or data archives. While there are reference specialists for specific topics or surveys, there are also general reference staff in data service who, for instance, run the helpdesk or data service hotline. From the people who work in general data service, two were chosen as initial participants in the study. Both of them had been working in general data service for several years, the first interviewee for

twelve years and the second interviewee for about five years. It was expected that they should have a very broad view of all kinds of needs and problems that data users could have. This assumption proved to be true. Not only did both participants shed light on a whole spectrum of complexities in data seeking behaviour; they also described their practice of directing users to specialists or other people, depending on their enquiries. This information directly informed the theoretical sampling. It underscored the need to interview reference staff who are specialists for only one or a few particular survey programmes.

Again, when deciding on the first interviewee from specific data services, a particularly data rich case was aspired. The ideal candidate, in this case, was someone who has a very long experience with this kind of work. Fortunately, the next reference person that agreed to participate was someone who had been working in data service for over 25 years. As he explained himself, he started on this job when there was no such thing as downloadable datasets but only data on magnetic tapes that had to be mail ordered. This interviewee is responsible for data curation and distribution for an international longitudinal survey, covering a very broad range of political and societal topics. He offered insight into a multitude of user requests and the problems and needs that users had. A key take-away from this interview was something that had already been indicated by the first two participants: Users' data literacy and experience with data analysis seems to be a key factor in successful data seeking. The second finding that stood out in this interview was that users repeatedly had difficulties with data documentation, no matter how well and comprehensive the documentation was. Also, this interviewee raised awareness for the situation of shared responsibilities in producing and curating a large survey programme.

The fourth participant was yet another reference person with promisingly broad impact, but in another sense. He had far less experience than the third interviewee, only about three years. But he is responsible for a very well-known and frequently used survey programme and heads a whole team of people who are occupied with curation, distribution and support for the data from this survey. The key take-away from this interview was that people who work with survey data learn about them early on in their education. This was an assumption that already the first interviewee had made. But the fourth participant was able to explain in detail, how students and researchers come to know certain datasets and thus find data of interest. This information was very much to the point of the question on how secondary

data users are seeking data. However, it seemed necessary at this point to also interview at least two reference persons who are responsible for more than just one survey programme to determine how generalizable the statements of the third and fourth interviewee on the users of specific large survey programmes could be.

The next interview was conducted with a reference person who had been working in data service for about 24 years. From the beginning she was responsible for multiple study collections. At the time of the interview she was working for a large international longitudinal survey programme as well as some less prominent international surveys. This interviewee was the one who drew attention to a phenomenon that would become a core concept of the grounded theory: dataset communities. She confirmed much of what was said in the interviews before about how people come to know survey data during education and how their seeking for data is influenced by their experience and data literacy. Her emphasising of the community terminology revealed in retrospect, that this concept had been there in the third and fourth interview already, maybe even in the second one. The fifth interview was the longest of all six interviews, and this participant was able to contribute crucial details to the theory of problem solving by community involvement.

As it had been decided before, one more reference person of multiple surveys was interviewed next. This person was responsible for the curation and distribution of several national survey programmes and was in this position for about six years at the time of the interview. Basically, the key findings from the earlier interviews could be confirmed in this interview. In fact, a lot of statements were repetitive with regard to the earlier interviews, in particular: the trouble users have with documentation; the influence of users' seniority or experience on success in finding what they need; the very basic requests of inexperienced users; the specific requests of experienced users. Even though the interview was steered in a way to learn more about the role of dataset communities in data seeking behaviour, the participant did not contribute particularly new aspects on this topic.

At this point it appeared that no more data was needed to describe the already very comprehensive findings. The analysis that had proceeded alongside the data gathering had sufficiently revealed specific aspects of data seeking behaviour that have not been described in information behaviour research before. The analysis has yielded several key findings that

will be used to inform the quantitative exemplification of the theory (see subchapter "C.3.2.1. Key Findings and Hypotheses"). There are even some other interesting findings that would be to extensive to follow-up on at this point and thus cannot be included in the quantitative study (see subchapter "C.3.2.2. Other Findings"). On these grounds, it was decided to end the qualitative data collection after the sixth interview.

2.4 Coding and Analysis Using Constant Comparative Method

As mentioned above (subchapter C.1), grounded theory development proceeds by simultaneous data collection and analysis. Subchapters C.2.4.1 and C.2.4.2 describe how simultaneous coding and analysis were done using constant comparative method in the present study.

2.4.1 Open Coding and Focused Coding

Initial coding of the first interview began as soon as the transcript was finished and before the second interview was conducted. The same procedure was applied to the second and third interview. The codes were phrased using gerunds as it is recommended by Kathy Charmaz and others (Charmaz 2014). This way of creating very active codes proved to be beneficial in analysis, in that it helped to keep the data users' perspective, even though they were not the interviewees.

The statements made by the interviewed intermediaries were always viewed from the users' perspective and the codes were phrased accordingly. For example, interviewees indicated that they sometimes recommended alternative datasets to users who had been asking for specific datasets. The respective initial code was *being referred to alternative datasets* and not *referring to alternative datasets* (cf. codes depicted in Figure 5).

Around the beginning of initial coding of the fourth interview, the number of initial codes had grown to 348 (Annex 4). This large number of codes had become hard to handle. In particular, it became difficult to discriminate between some of the codes. For instance, the code *being referred to alternative datasets* was very similar to *learning about alternative datasets*. The same is true for the codes *coming from different disciplines* and *coming from disciplines other than the social sciences.* Another example is the similarity between the codes *asking for facts* and *asking for results*. To discriminate between certain codes, explanatory notes were written. Following the grounded theory approach of constant

comparison, some codes were interrelated using the atlas.ti functionality of creating links between codes. Meanwhile, a total of five memos had been written (Annexes 7–11), which were aimed at discussing the impact and potential interrelation of the most interesting codes or their possible roots and contexts. For example, some initial codes surround the phenomenon that data users ask for information that they could easily have found on the web. A memo was created that detailed possible reasons for this behaviour (Annex 8).

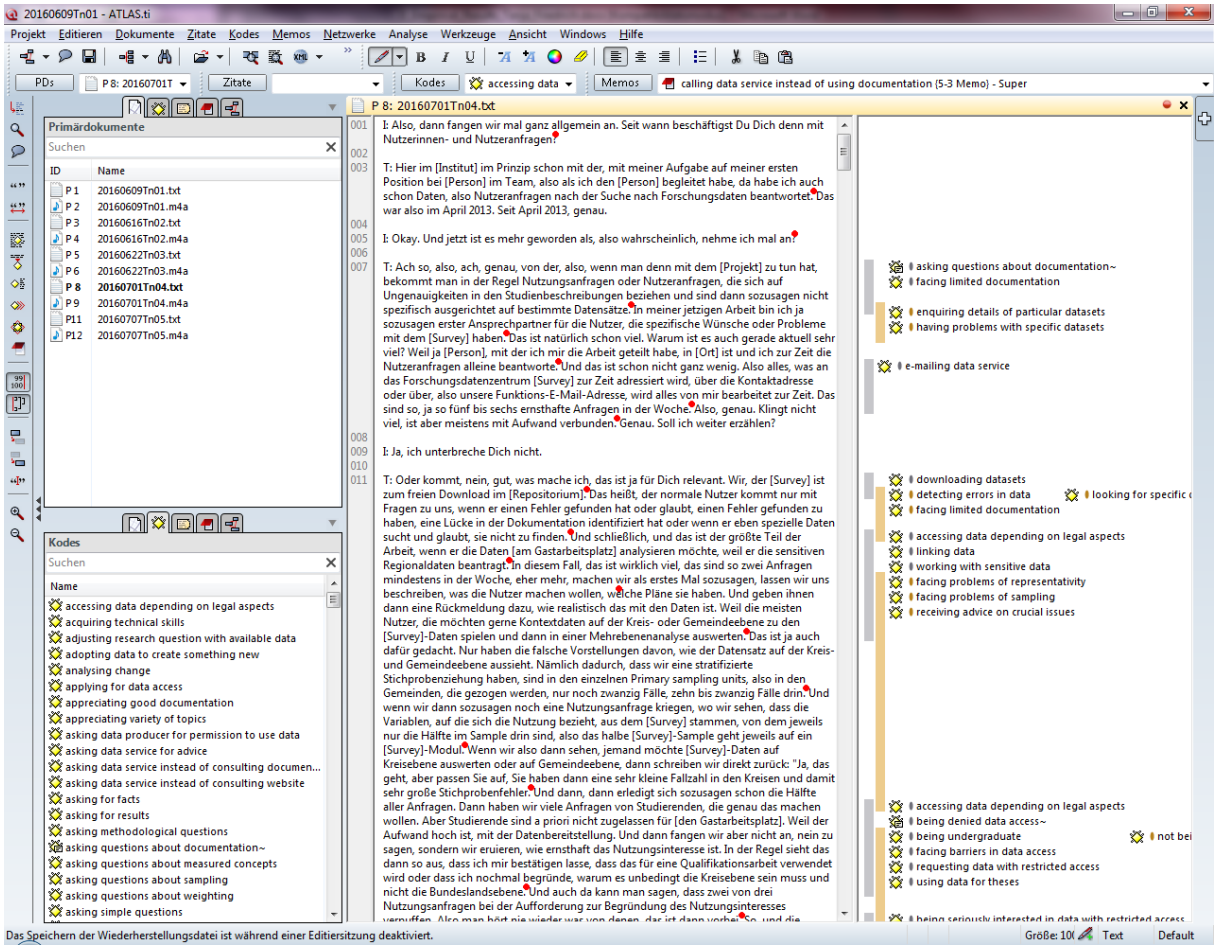


Figure 5 Initial coding of interview no. 4 (screenshot from atlas.ti)

Scrutinizing the codes in that way led to the creation of eight code families in an attempt to give more structure to the large collection of codes (Annex 5). The code families were:

- Being diversely skilled
- Being influenced by external factors
- Employing different styles of request

- Facing problems and barriers
- Having a certain affiliation, profession or education
- Interacting with data service
- Knowing and learning about data
- Satisfying a particular goal

Meanwhile, the theoretical sampling as well as the ongoing interviews were influenced by the analytic impact of scrutinizing and comparing the codes. Around the beginning of the coding of the fourth interview, it became clearer, which codes or code families were the most promising in terms of developing a grounded theory of information seeking behaviour of secondary survey data users. For example, the code family *Interacting with data service* did not seem to have enough analytical power to be pursued intensively. The reason why this code family emerged and also comprised the most codes (71 out of 348) was simply that the interviewees talked about their experiences of working in data service. The families *Satisfying a particular goal* or *Being influenced by external factors* seemed to have much more analytical power with regard to the research question.

On these grounds, the analysis went on from initial coding to focused coding. The three already coded interviews were re-coded with the new focused codes. All further interviews were directly coded with focused codes. While the initial codes had been more descriptive in nature, the focused codes were phrased more conceptually, resulting from asking what was implied or revealed by the initial codes (Charmaz 2014) with regard to the research question and the interviews that had been already conducted. The memos and code families helped this development. The resulting list contained 182 focused codes (Annex 6).

To reach more conceptual structure, all focused codes were scrutinized and compared with regard to commonalities and differences right from the beginning and a categorizing prefix was added to every code. For example, the codes *coming from another discipline* and *working for governmental institutions* were given BACKGROUND as a prefix: *BACKGROUND_coming from another discipline* and *BACKGROUND_working for governmental institutions* (cf. codes depicted in Figure 6).

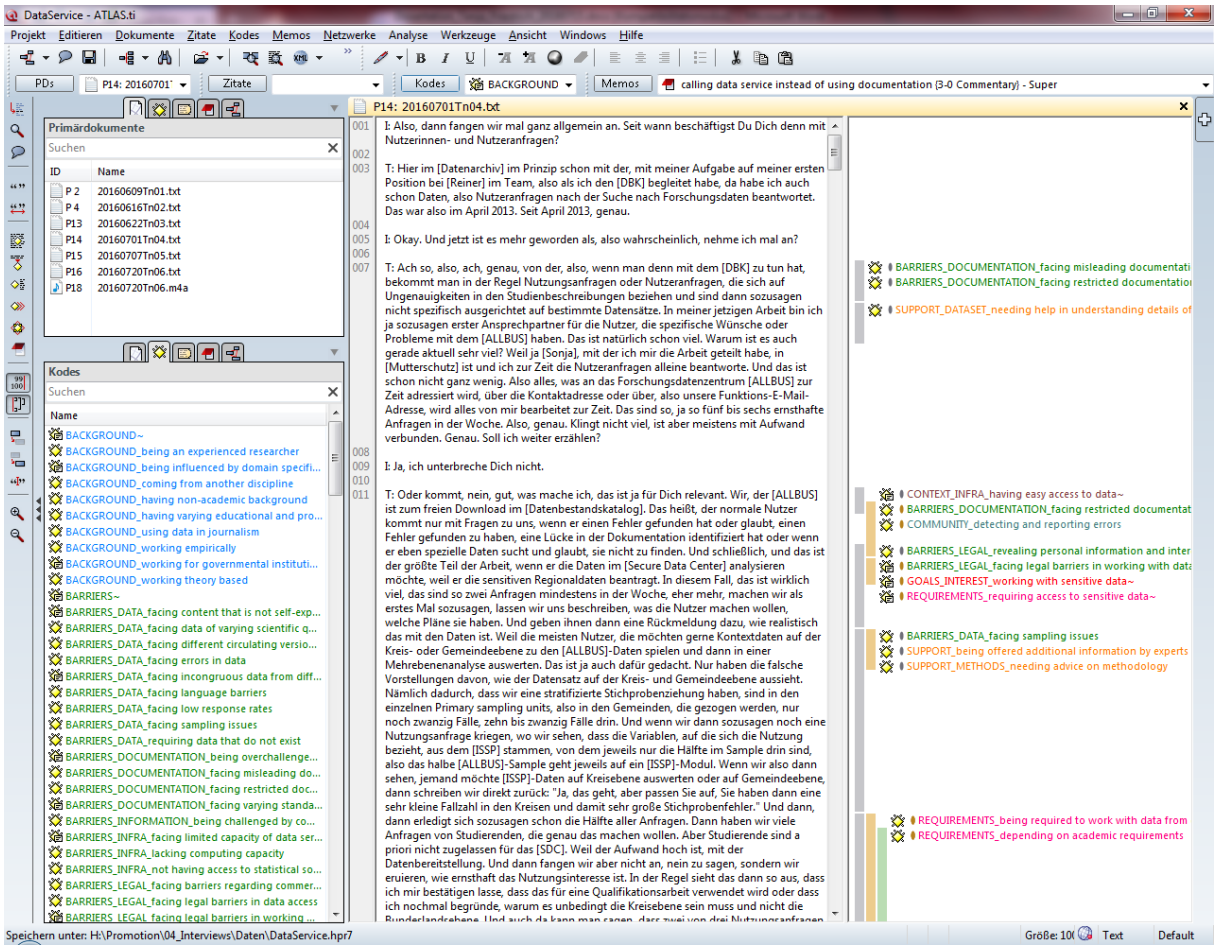


Figure 6 Focused coding of interview no. 4 (screenshot from atlas.ti)

Some of these prefixes grew to contain significantly more codes than others. Those were specified further by sub-prefixes, such as:

- GOALS_INTEREST
- GOALS_METHOD
- GOALS_NEED
- GOALS_SUCCESS
- GOALS_TASK
- GOALS_UNCLEAR

The categorizing prefixes were treated as tentative categories of the grounded theory of information seeking behaviour of secondary survey data users. In the end, eight[14] categories were identified that are briefly explained in the following list with regard to the related data:

- BACKGROUND

  This category refers to the professional as well as educational background that users of survey data may have. All interviewees indicated that users came from different professions and by no means only from academics. There are also non-academics such as journalists or school teachers who are interested in survey data. With regard to academics, the interviewees indicated that researchers from different disciplines and with all levels of experience in data analysis were among the users. It became clear from the interviewees' accounts that usually, the professional or educational situation determines requirements and constraints in data seeking or use. The category BACKGROUND comprises nine codes. In all interviews, a total of 58 quotes were coded with BACKGROUND codes.[15] The category BACKGROUND turned out to be an important category in the qualitative study.

- BARRIERS

  This category refers to problems and barriers that may occur when seeking data, in data access, and in data use. The category BARRIERS comprises 24 codes. During coding, seven subcategories were identified:

  BARRIERS_DATA: barriers with regard to the data(set), such as errors in the data;

  BARRIERS_DOCUMENTATION: barriers arising in the context of data documentation, for example, from restricted documentation;

  BARRIERS_INFORMATION: barriers resulting from overly complex website information;

  BARRIERS_INFRA: barriers concerning lack of infrastructure such as support services;

[14] Initially, there was a ninth category CONTEXT that turned out to be the least used category. As it turns out, this concept is too broad and ambiguous, since all influential factors can be interpreted as context factors. Potentially relevant quotes were rather coded with more specific categories such as BACKGROUND, REQUIREMENTS, BARRIERS_INFRASTRUCTURE, or COMMUNITY.

[15] The number of quotes coded with a code does not equal the overall occurrence of the phenomenon in the data. For example, when interviewees kept indicating that experienced social science researchers were looking for data, the respective code "BACKGROUND_being an experienced researcher" was treated as saturated from a certain point on. In later interviews it might have been only used again if a new aspect of the phenomenon had emerged that would help to capture the phenomenon more holistically.

BARRIERS_LEGAL: legal barriers with regard to access restriction;
BARRIERS_SEEKING: barriers that occur in the process of seeking such as failing online services;
BARRIERS_SKILL: barriers that result from a lack of skill, when trying to figure out the content of a complex dataset.
In all interviews, a total of 152 quotes were coded with BARRIERS codes. Even though this category was used a lot in coding, its influence in theory-building was not equally strong as the influence of the categories BACKGROUND or GOALS. The high frequency of quotes on BARRIERS results from the fact that the interviewees reported from their experiences with user requests, which typically arise when people encounter barriers. For the general description of information seeking behaviour, barriers are a relevant factor, but not a constitutive one.

- COMMUNITY
  This is the core category that emerged from the qualitative study. It serves as a container for various codes that suggest that community involvement influences or even determines data seeking. Surrounding a survey or a collection of datasets, there are communities that are made up from people who play any role in the preparation, distribution, finding, and use of these datasets. The category COMMUNITY comprises 16 codes. The codes represent various community activities, such as being referred to experts or data service, contributing to documentation or finding and fixing errors in data, sharing data informally, or even networking to get access to restricted data.
  In all interviews, a total of 105 quotes were coded with COMMUNITY codes.
- GOALS
  This category is a very strong category that turned out to play an important role in the qualitative study. This is not surprising given the definition of information seeking behaviour as “goal-oriented problem solving” (Wilson). The category GOALS comprises 43 codes, more than any other category. During coding, six subcategories were identified:
  GOALS_INTEREST: goals that reveal a specific research interest, such as studying change or working comparatively; also: goals that indicate the type of information that a user is interested in, such as results rather than data.

GOALS_METHOD: the goal to apply a specific method of analysis to the data, such as multivariate analyses or georeferencing.

GOALS_NEED: the underlying information need, such as the need to know how many people believe in a certain concept.

GOALS_SUCCESS: the ultimate goal of being successful in research, for example by doing original research, working with most recent data, following research trends, or getting published.

GOALS_TASK: the actual task that the data is needed for to reach a certain goal, such as learning to work with data, using a dataset for methodological exercise, using a dataset in a research project, doing a replication, sorting out a subject of research by looking what data are available, measuring a concept of interest, or reusing a dataset's measures to gather own data.

GOALS_UNCLEAR: goals with an unclear intention, such as the goal to just collect datasets, independently of need; or such as preferring large or popular surveys or, somewhat inversely, preferring unique datasets.

In all interviews, GOALS codes were assigned 209 times in total.

- REQUIREMENTS

  This category turned out to be one of the most influential categories, along with the categories GOALS and BACKGROUND. The category is central to the grounded theory in that data seeking as well as data use is always subject to various requirements, constraints, and dependencies. REQUIREMENTS can be seen as a context factor and are closely linked to GOALS (there is an overlap between the category REQUIREMENTS and the category GOALS_NEED) or arise from GOALS (such as data quality requirements if high end research is the goal). Also, REQUIREMENTS are influenced or even determined by professional and educational BACKGROUND. The category comprises 11 codes that refer to academic or educational requirements and requirements arising from the topic or kind of research that is practised. Quotes coded with "REQUIREMENTS_requesting recommended or stipulated datasets" indicate that REQUIREMENTS also arise from COMMUNITIES. In all interviews, REQUIREMENTS codes were assigned 71 times in total.

- SEEKING

  This category refers to seeking and finding data in various ways as well as associated problems or challenges. Obviously, it is a key category of the grounded theory of information seeking behaviour of secondary survey data users. Strong codes in this category indicate the importance of informal ways of information seeking, for example, the codes “SEEKING_SOURCE_consulting intermediaries”, “SEEKING_SOURCE_learning about data in academic or educational contexts”, and “SEEKING_SOURCE_receiving biased data advertisement”. These and other codes are strongly related to the core category COMMUNITY. The category comprises 28 codes. During coding, five subcategories were identified:

  SEEKING_CITATIONS: refers to instances where DOI citations are used to access data, which indicates that users are chaining from literature; and the fact that people use frequently cited data.

  SEEKING_DOCUMENTATION: making use of documentation in the process of data seeking and having to deal with comprehensive or restricted documentation.

  SEEKING_RELEVANCE: refers to relevance judgement when seeking data, for example with regard to subject relevance in general or with regard to measures taken to determine relevance.

  SEEKING_SEARCHING: refers to the concept of information searching as a narrower concept of information seeking (cf. Wilson); for example, users are searching data catalogues, or scanning datasets for relevance, or searching known datasets or variables within known datasets.

  SEEKING_SOURCE: refers to the source that informs users about the existence of surveys or datasets (literature, media, advertisement, teachers, peers) or that is used to search for data (catalogues, search engines); also, sources that give further information or support when looking for or working with data (intermediaries such as data service); and ways of using these sources.

  SEEKING codes were assigned 93 times in total.

- SKILLS

  Personal SKILLS were revealed by interviewees to be a relevant factor in seeking, finding, and using data. This category refers to users’ skills with regard to finding and using data, understanding and using data documentation, applying statistics, and

legal knowledge. To determine the influence of SKILLS on data seeking behaviour, it proved helpful to discriminate the codes according to their positive or negative influence (SKILLS_POS and SKILLS_NEG). Positive factors with regard to data seeking behaviour are good skills in finding data; the ability to find errors in datasets; and having empirical and statistical skills (which is part of survey data literacy). Negative factors are, for example, being oblivious to documentation or to errors in data; lacking knowledge in statistics (or not being survey data literate in general); not being able to understand documentation or to read data; lack of knowledge in statistical software. From the 17 codes in this category, 12 were identified as SKILLS_NEG and 3 as SKILLS_POS. The remaining 2 codes are neutral with regard to their influence on data seeking behaviour: “SKILLS_employing simple statistics” and “SKILLS_working with noncomplex datasets”. SKILLS codes were assigned 72 times in total.

- SUPPORT

  With 179 quotes that have been assigned SUPPORT codes, this category is one of the most represented categories in this analysis. Given that the interviewees were not people who are looking for data but intermediaries who support them in finding data, it is not surprising that the interviews produced many codes on SUPPORT issues. This has to be considered when assessing the impact of the category. Instances of SUPPORT only occur where users need assistance, and these instances are those that dominated the accounts in these interviews. The category is relevant for the theory of data seeking behaviour, but it is not a key category. It informs the analysis about the use cases that require support.

  The category of SUPPORT includes all requests made to data service personnel, but also to other people such as colleagues, supervisors etc. During coding 8 subcategories were identified:

  SUPPORT_ANALYSIS: needing, being offered, or making use of support in data analysis;

  SUPPORT_DATASET: being offered useful tools together with a dataset or alternative, pre-released, or specifically processed data; needing help to work with a specific dataset;

  SUPPORT_DOCUMENTATION: refers to advice on or instead of documentation;

  SUPPORT_METHODS: training or advice on methods;

SUPPORT_PERSONAL: choosing or being offered personal contact for support;

SUPPORT_RESEARCH: needing help with research design;

SUPPORT_SEEKING: receiving support or advice to seek or search for datasets;

SUPPORT_TECHNICAL: needing technical data service.

Eight codes could not be assigned to one of the subcategories. Most of them indicate referrals of some kind, for example, referrals to literature on the data.

The categories that have been created through the process of focused coding relate to the areas of exploration presented in B.3 and to the interview topics from the interview guide (C.2.1) as it is depicted in Table 2. The data that have been coded with these categories provide information on these areas and topics.

Table 2 Areas of exploration, interview topics, and categories from focused coding

Area of exploration	Interview topic	Category
The users' educational/professional background	Educational/professional background of data users	BACKGROUND
The users' research experience and data literacy	Survey data literacy	SKILLS
Information sources and channels; the role of intermediaries and information technology	Users seeking support	SEEKING; COMMUNITY
Goals, needs and purposes of users	Tasks and goals of users	GOALS
Requirements when looking for data	Relevance criteria and data quality	REQUIREMENTS
Requirements when looking for data	Research trends	REQUIREMENTS
The role of documentation	Information and documentation	SEEKING; BARRIERS; SUPPORT
Barriers and problems when looking for data	Problems and barriers	BARRIERS

2.4.2 Memo-Writing and Theory-Building

Memo-writing is a crucial analytic step in grounded theory methodology that allows for informal reflections on data and coding (Charmaz 2014). It helps the development of concepts (or theoretical codes that are based on abstract concepts (Herring 2013)) and relationships between them and furthers theory building. For the present study, the

technique of memo-writing was employed during initial (open) coding as well as during focused coding. The resulting memos helped the development, comparison, and clarification of codes during initial coding and of tentative categories during focused coding.

As indicated above, five memos were written during initial coding (Annexes 7–11). These memos were mainly used to investigate the impact and potential interrelation of the created codes. The main goal was to identify relations or structures among the codes that would help to sharpen the focus with regard to broader theoretical categories. The writing of these first memos contributed to the transition from open coding to focused coding by identifying relationships and common grounds of initial codes. The most useful and theoretically solidified memos were written during focused coding and after the coding was finished (Annexes 12–15). In the end, nine memos had been written, some of them connected with codes, some of them connected with quotes, and others independently from specific codes or text. One of the most interesting memos was written after the fifth interview. The interviewee had helped identify and name the concept of dataset communities which had been present but unvoiced in earlier interviews. Based on this insight that was fostered by the fifth interviewee, the memo “dataset communities” had been written (Annex 9). In the end it became sufficiently clear that dataset communities are a significant factor in data users’ information seeking behaviour. With the research question in mind, a final memo on “the theory of problem solving by community involvement” (Annex 15) was written.

3. Results

3.1 Key Codes and Categories

When assessing the tentative categories together with their related data, it became clear that two of them had less analytic power (cf. Charmaz 2014) than the other six categories. These were the categories SUPPORT and BARRIERS. The category SUPPORT seems less relevant or even redundant when taking into account that the interviewees are professionals whose core task it is to give support to people who are looking for or using datasets. The same bias seems to be relevant for the category BARRIERS, because users who do not experience barriers do not seek help from reference staff. In turn, the most promising categories seemed to be BACKGROUND, COMMUNITY, GOALS, REQUIREMENTS, SEEKING,

and SKILLS. These categories as well as tentative relationships between them serve as cornerstones of the grounded theory of information seeking behaviour of secondary survey data users that was developed on the grounds of the collected interview data. The memo "problem solving by community involvement" (Annex 15) was a first verbal version of the grounded theory. The following diagram (Figure 7) depicts this theory in a schematic way:

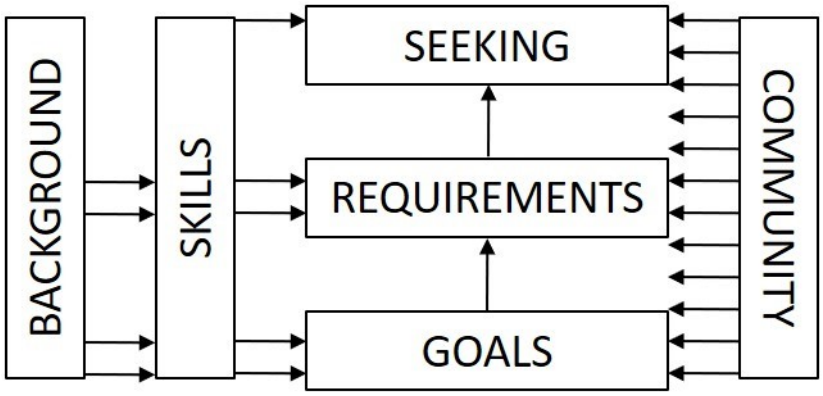


Figure 7 Schematic diagram of the theory of problem solving by community involvement

The core categories are depicted and arranged in six boxes. The interrelation between the boxes (categories) is depicted by arrows. The arrows that originate from BACKGROUND and SKILLS on the left depict the influence of these categories on GOALS, REQUIREMENTS, and SEEKING. There is only one arrow from SKILLS to SEEKING signalling that the direct influence of SKILLS (in the sense of data literacy) on SEEKING is less strong than the direct influence of SKILLS on GOALS and REQUIREMENTS. The influence of the BACKGROUND on SEEKING is only seen as an indirect influence, whereas there is direct influence of BACKGROUND on SKILLS. The arrows from GOALS to REQUIREMENTS to SEEKING imply a sequential arrangement. This means that, according to the theory, goals determine requirements and requirements determine seeking. The core category COMMUNITY on the right provides the background for all the other categories and the processes and dependencies that are depicted by the arrows between these categories. This is why the arrows that originate from the COMMUNITY category point towards the whole arrangement of the other categories and relationships.

To explain the nature of these influences and dependencies in a deeper way and to prepare the introduction of the comprehensive theory, the following paragraphs are dedicated to a

more detailed assessment and description of these categories and the relationships between them. To do so, significant quotes of the participants are provided that show how the categories and relationships between them are rooted in the collected data. By presenting and interpreting quotes given by the interviewees, this analysis explains the cornerstones and constitutive processes of the theory of problem solving by community involvement.

BACKGROUND and SKILLS

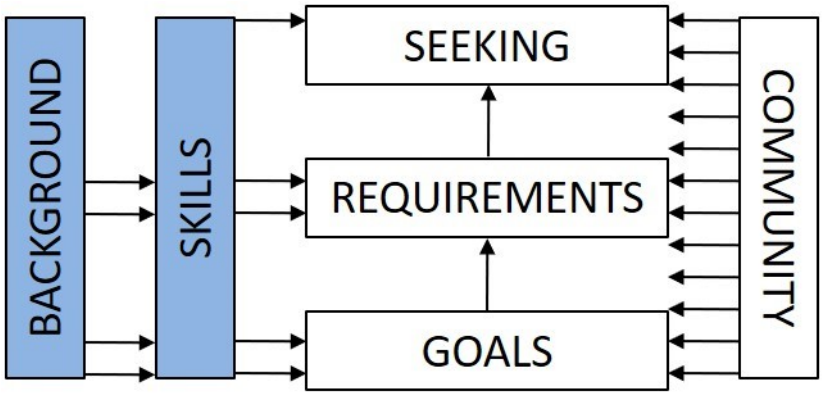


Figure 8 Background and skills

As expected at the outset of the study (B.3 Areas of Exploration), the users' educational and professional background seems to play a fundamental role in different aspects of data seeking behaviour. The users' background was addressed at some point during every interview, most of the times in association with questions of skill and data literacy. When asked about the general data users' expertise, experience, and background, the interviewees described a broad spectrum of professional experience, academic seniority, data literacy, data analysis skills, and subject interests. In general, clients of data services include university students who work on assignments as well as advanced researchers and experienced professors, but also people from outside academia: "Politicians, journalists, students, the general public; that is the user group of [survey name]."[16] (In05, 627)[17] Some interviewees also named school teachers as their clients. One of the major international

[16] All quotes that are presented here were translated from the German original.
[17] Quotes from the interviews are referenced by a combined abbreviation of 'In' for 'interview' and the chronological number of the interview and the line number from the transcript.

surveys even produces material for school teachers based on their data: “Plus, there is this education web for secondary, for *Abitur* or *Gymnasium* level, I believe. In seven or eight languages, meanwhile. So students can discuss [survey] topics internationally.” (In05, 634) The different backgrounds lead to diverse skill sets. Some of the users are experts in survey data analysis, others do not even know what a survey dataset looks like. Some users from outside academia or from other academic fields are even surprised that they are not offered statistics or results of any other kind, but have to do their own analyses with the data: „Our understanding of data, what we are in charge of, are micro-level data. But to many users’ understanding, data are aggregate data, that is to say survey results, tables.” (In03, 15) These outsiders sometimes hear or read about politically relevant surveys (such as surveys on voting behaviour) on TV or in the newspaper and then try to find information on these surveys: „The results are always published in the press. Well, because it is a political survey, and that makes it highly visible.” (In03, 127)

BACKGROUND and SKILLS influence GOALS and REQUIREMENTS

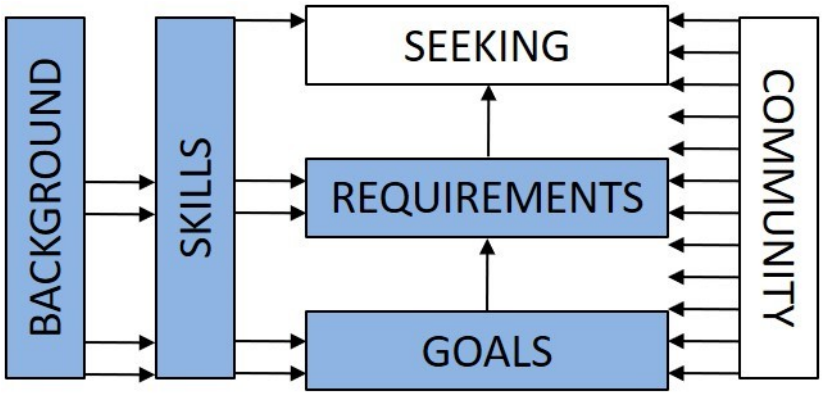


Figure 9 Background and skills influence requirements and goals

The background and data literacy of people who want to work with survey data determine their information goals as well as their requirements when looking for data. Common requests of users that come from outside survey research show their interest in results instead of data quite blatantly: “I can tell it by their requests that are anything but theoretically permeated, but simple, plain. Yes, in part, they want to retrieve facts. Along the

lines of: ‘Is it true that we, that the social divide here in Germany has grown over the last ten years?’” (In02, 147) These users do not need datasets, they only need some specific results based on survey data: “Many PhD students, coming from edge areas of the social sciences, who need that as an add-on to their actual work, in a way. Descriptive attachment. Geographers, theologists; those who just want to adduce some countings, for whom this isn’t the focus of their attention.” (In04, 199) For these users, some of the large survey programmes include tables on key issues in their variable reports. And sometimes, very experienced data service staff can answer specific fact retrieval questions without even looking into the datasets: “Often, there are these requests from journalists. Actually, this happens once a week. Often combined with a request to do data analysis as well. So, they are looking for information, descriptive information on a specific topic. What’s the most recent one that I had? Values in Germany. Regional differences concerning values in Germany. This topic is addressed quite frequently. And there I can tell them from the outset, there are no differences. I have processed this so many times in the past that I can say: No, you don’t need to look any further.” (In04, 68)

The more experienced users know exactly what they want to do and what data they need for it. Others, usually students or other novice users, do not even know what topic or phenomenon they want to investigate: “There are these requests: ‘I have to write this assignment, do you have any interesting topics?’” (In05, 96)

At one point or another, every interviewee indicated that a user’s experience with data analysis clearly was an influencing factor in data seeking. For instance, one interviewee explained: “I really do believe that experience plays a part in that as well. I mean that students, in particular, who are approaching a big project like that for the first time or maybe even for the second time, maybe they don’t even know full well what they are actually looking for.” (In06, 647)

GOALS, REQUIREMENTS and SEEKING

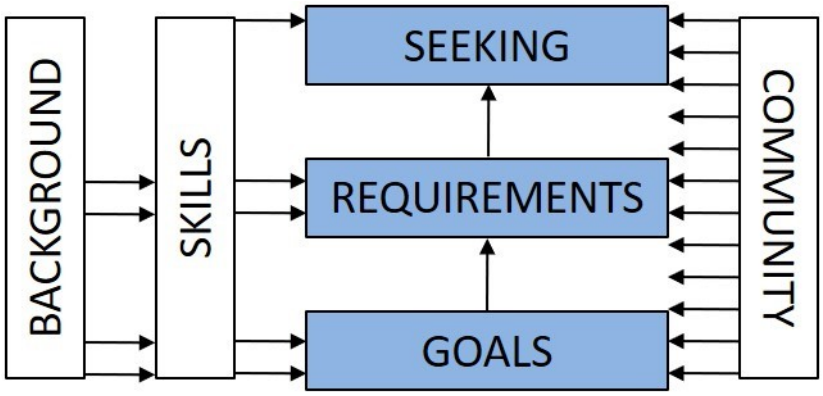


Figure 10 Goals, requirements, and seeking

Regardless of professional or educational background, users' goals seem to determine their requirements, which then influence data seeking practices. Apart from subject relevance, data users have a couple of other important requirements, according to the interviewees. These requirements depend on the goals that they want to reach. For example, journalists or undergraduate students might look for a dataset that allows them to find answers on their topic of interest with the least possible effort: "Like, does the data answer this kind of question? Or how much do I have to do to get an answer with the help of this data?" (In02, 83) For journalists in particular, time constraints are another requirement: "Traditionally, that is, or in those cases that we have had so far, it is the journalists who, most of the time, need these things on relatively short notice." (In06, 419) A requirement that occurs with more ambitious goals is that data contain sensitive information: "[…] if they want to analyse data in a safe environment, because they apply for sensitive spatial data." (In04, 24) In general, specific methodological interests can lead to respective requirements: "Well, I always see a mixture, subject questions, subject interest regarding the topic, combined with methodological interest or statistically methodological interest." (In02, 288) Finally, data quality seems to be a relevant factor, at least for advanced researchers: "It has to be easily reproducible. So, clear variable labels. Very clear codes, so that, which are labelled properly, without shortages. A reasonable concept for missing values becomes more important as well, because missing values have become very professionalised and more differentiated." (In04, 672)

Generally, it could be assumed that subject relevance should be the most important influencing requirement that data users can have and that subject relevance was the most important factor in data seeking. Not surprisingly, when asked about users' inquiries, the first two interviewees indicated that a typical request was for data on a particular research topic. For example, a student would ask: "I am working on a thesis on topic XY, what data can you offer me?" (In01, 100); or a researcher from a larger research group would say: "We need empirical material on topic X."(In02, 65) With these initial requests, some users turn directly to data services; others appear to perform more or less extensive online searching before writing or calling for support. For instance, users who have already found interesting data would ask: "I could probably use this survey for analysis. Are there any other sources that I might have missed?" (In02, 92)

It is interesting to note here that trying to find data by performing keyword searches on the web, in data repositories, or data catalogues is often unsuccessful. There seem to be different reasons for this problem, one of them being that standardized subject indexing for survey data does not exist (Friedrich and Siegers 2016). Another reason is that in the social sciences, the same concept can be defined and studied in multiple ways by means of operationalisation (Friedrich and Siegers 2016) (cf. memo "Concepts and indicators in secondary analysis", Annex 11). The situation is such that there is rarely just one way to survey a certain issue. As a result, different population surveys ask their respondents different questions even though they are investigating the same topic. Secondary users of these surveys have to examine the questions closely to judge their relevance: "The survey questionnaires are the best starting point, of course, since they contain the exact wording of the questions. Because sometimes, this is all that matters; a survey question can be aimed at the same topic but be different in phrasing. And just like that, two datasets turn out to be incomparable." (In01, 436) This means that, even if there is a dataset on the concept in question, it may be measured in a way that is not useful for the secondary user's purpose. One interviewee explained the difficulty in finding exact matches when looking for data on a specific topic: "Then I go and enter it in the search field, get a result, and look at it. And here and there I look at the datasets and I see: Yes, almost. A close one. And the next one is also a miss. And somehow you don't get a fitting result that is an exact match of what you are looking for. They are all only nearly there, but not quite." (In01, 508)

Given these difficulties in finding appropriate data by conceptual keyword searching, it is no surprise that, however common these initial search requests may be, they are not necessarily the most important requests in data seeking behaviour. They may not even be the most frequent ones. From early on in the field work it was apparent that secondary users of survey data are not just interested in data that are thematically relevant for their research. To a greater degree, researchers apparently come from a position where they already know about the datasets that they could possibly work with: “Well I, you do realize whether people come up with a research topic first and then start looking for data or if they do it the other way around. Like at first, they take a look at what data is there, what information is available. And then they come up with a research topic. Of course, this is much easier than searching over and over again without finding anything.” (In01, 513)

From early on in the interviews one important aspect of survey data research practice became more and more clear: Instead of being entirely open with regard to reusable datasets on specific topics, researchers seem to have a strong interest in working with data that they already know. Specifically, many researchers intend to work with high quality datasets from large survey programmes that are well known in the research community. At this point, it seems that goals, requirements, and seeking with respect to survey data are not only influenced, but shaped by the research community that the users of this type of data belong to.

The survey data COMMUNITY

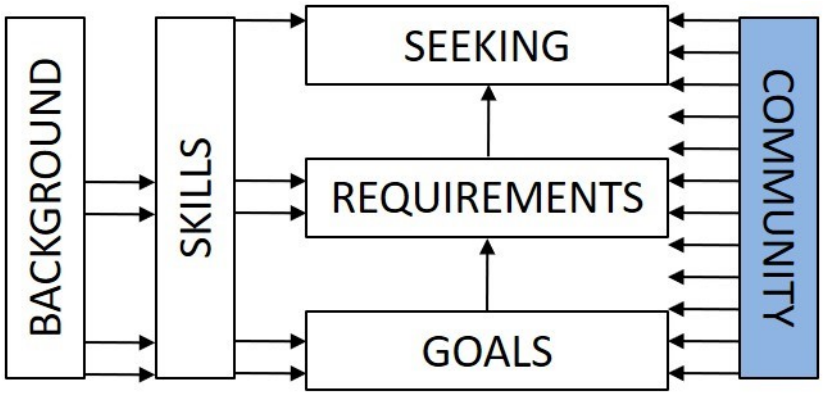


Figure 11 The survey data community

At a certain point during the interviewing process, around the fourth and fifth interview, it had become sufficiently clear that for every large scale survey programme, there is a whole circle or community of people who are intensively and repeatedly working for and with the datasets from this particular survey. The interviewees revealed and confirmed the idea that there were different groups of people who are concerned with particular datasets. Interviewees indicated that those communities were not only made up by the users, but by a whole range of people who each fulfil different roles in the production, curation, distribution, and re-use of these datasets. One interviewee pointed out the advantage of division of labour in survey data research was to being able to "divide up chores in a nice way", explaining that, for example, "for aggregated results, go to [principal investigator] and microdata are available over here." (In03, 284) At one point, an interviewee who is responsible for the curation of a large longitudinal, international survey programme established that "surrounding this survey, there is a community that shares a common investment" (In05, 230). This participant explained that this community consisted of a whole range of people taking over different roles with regard to producing, improving, distributing and using data from a specific survey programme.

Communities in survey research have, use and provide particular knowledge on certain survey data programmes or groups of survey data programmes. These large survey programmes produce datasets that are intended and designed for secondary use and thus receive extensive financial funding. Compared to data from smaller surveys, these data usually have extended (added value) documentation and the programmes offer special data services, such as newsletters, conferences, and other events: "There are respective workshops, like 'meet the data' for [survey] or [survey]." (In01, 567) Large survey programmes are advertised prominently and often carry a certain prestige that makes many researchers want to work with them. The second participant in the interviews explained: "I do think that there's something like prestigious data or flagship surveys. Like, whether you look at the very well documented [survey name] that is requested frequently, eagerly, and is used in many publications. Or the datasets from [survey name] that are eagerly requested even by the international community, too. People use them a lot for their work." (In02, 603) These datasets stem from longitudinal surveys that produce new datasets in chronological

waves, for example, every year or every other year. There are a couple of very well used survey programmes that offer cross-national data as well as programmes that survey the same individuals over time (so-called panel surveys): “Well, I know that for some years now, panel surveys are very much in vogue.” (In01, 273) These studies are especially valuable for comparative analyses and are therefore very popular. Conveniently, large survey programmes commonly cover a broad spectrum of research topics. And maybe most importantly, they take the work of data collection out of the researchers’ hands and invest reliably in high and consistent data quality. Not surprisingly, data quality is an important criterion when looking for data: “The most used surveys are of very high quality, one would have to say that. Meaning, you don’t have some messy survey that is used a lot. That wouldn’t work.” (In01, 573)

The second interviewee suggested, that the extensive documentation that is provided with the datasets of these surveys is another important reason why researchers use these datasets a lot: “Many of them, I’d say just the popular ones like [survey name], [survey name] and many others, [survey name], have great documentation. They have commendable documentation. And this is information that our users like, that is the impression that I have. So, they are very interested in using well documented survey data.” (In02, 326)

COMMUNITIES influence GOALS, REQUIREMENTS, and SEEKING

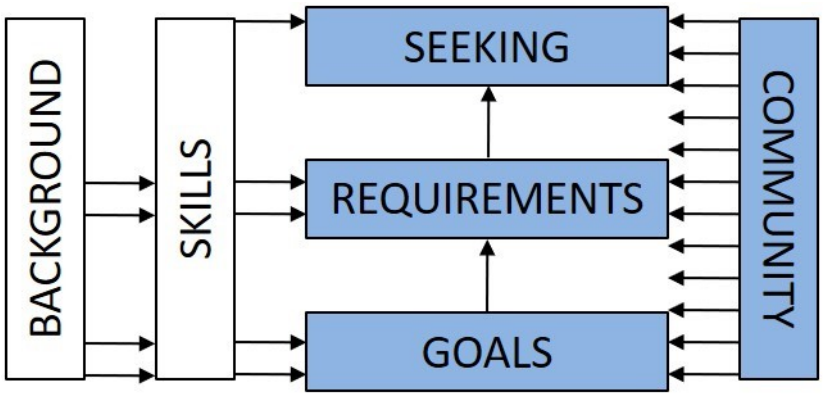


Figure 12 Community and goals

Already the first interviewee indicated that large survey programmes were in high demand among data archive clients: “Well, when a large survey is published, there is a certain run [on it], to some extent.” (In01, 9) People seem to be waiting for data from new waves of a particular survey to be released: “We just published the seventh wave of this survey. I have, one of the customers, he already ordered the third, fourth, fifth and sixth wave.” (In01, 530) Other interviewees confirmed later on that, in the case of longitudinal surveys, now and then users ask whether a new dataset has been published yet: “It is four months in advance that users start asking: When will the new [survey name] get published? When will [survey name] be available?” (In04, 514) Apparently, there are researchers who concentrate on these specific surveys, to which they always come back to, even though there might be alternative datasets on the same topic: “I always tell people, there is this other survey [survey name] on [topic]. Meanwhile, there are seven waves available, I believe. Actually quite good as well, maybe not that comprehensive. But no, people prefer this survey [survey name]” (In01, 551). This suggests that, even if users are presented with alternatives, they tend to cleave to data from familiar surveys. Following up on this indication, the next interviewees were specifically asked, whether there were users who repeatedly come back to the same data or surveys. In ways of theoretical sampling, the tracing of this phenomenon led to interviewing reference staff who are more in-depth specialists for only one or a few datasets (see subchapter "C.2.3 Initial and Theoretical Sampling"). They confirmed the initial impression that there are indeed users who prefer to work with datasets from a particular study: “I believe there are only few who we can interest in other survey programmes.” (In04, 313) To the contrary, people who have worked with data from a particular longitudinal survey programme in the past are interested in continuing their work with new data from this programme. This may also explain, why there is always a certain run on newly published datasets of large survey programmes (see above). Fittingly, all interviewees confirmed that many users express their need for recent data.

One interviewee who is responsible for the curation and archiving of a very well-known longitudinal survey tried to explain the success of this programme: “Well, the [survey] is a brand.” (In04, 251) Measurements of this particular survey are even being used as templates in the design of new surveys: “You just take the instrument from [survey name]. This is a big advantage, because you can just say: I measured education as it is done in [survey name].

[…] I measured political interest as it is done in [survey name]. […] It is done like this, it is very common. […] This is a standard dataset; it is the dataset of reference." (In04, 712) Referring to the survey as a "brand" points to the reputation that it apparently has in the community. Most interestingly, according to the interviewee, the continuing popularity and wide reuse of this dataset is not as strongly tied to its quality as one might think. In fact, while this survey has been around for several decades, it has well-known flaws and today, there are competing surveys on the market. The interviewee's guess on why it is still so popular is: "I believe, what is really important is that you, that [survey name] really is a brand. A well-established reference survey that is taken seriously; still taken seriously in spite of its weak points." (In04, 825) Apparently, not all researchers are applying the highest quality standards to their data: Interviewees have indicated that, as long as a dataset somehow does for them what they need, some users are inclined to ignore problems such as low sampling rates. One interviewee reported several such instances and how he always tries to explain to the users that reliable assertions cannot be made on the grounds of these data. The interviewee concluded rather resignedly: "And then I get the impression that they don't really want to hear this." (In03, 648)

Then again, one of the other interviewees who is also responsible for a large survey programme insisted on the importance of data quality and good documentation for creating a product that can compete with the other surveys that are out there: "There certainly is competition between survey programmes. If you have a lot of errors in your data collection and leave it like that, people will turn to [other survey]." (In05, 232) The interviewee concluded: "Users who turn to that survey programme are on the safe side." (In05, 275) Of course, there are other factors than just longevity and quality that make up for a survey's success. Sometimes it is more about the reputation of the people involved than of the survey itself or as the same interviewee put it: "They have renowned researchers on their boards. The surveys' boards truly are very important." (In05, 284)

The interviews suggest that there are diverse ways of how people come to know certain surveys. "For the most part, I think, through literature" (In01, 37), one participant indicated. Another interviewee agreed with this suggestion and added that users learned about surveys "at conferences, which is where they meet with regard to their research," (In05, 268) referring to more advanced researchers. For novice users like students, citations in

reference literature should be the more important source. One interviewee explained: “If you look at textbooks, textbooks on methodology for the social sciences, [survey] is always one of the mentioned data sources.” (In04, 851)

This suggests that the knowledge of and preference for certain survey programmes is initiated early on in an academic career. One interviewee indicated that an increase in students’ request was particularly noticeable: “Well, you just notice that people are confronted with data at a much younger age now, much earlier.” (In01, 121) According to the interviewees, senior researchers tend to introduce junior researchers to popular or frequently used datasets, for example when they are teaching statistics to students: “Every sociology student, also most of political science students, many students from economics, many from geography make their statistics education with [survey]. […] Every student knows about the [survey].” (In04, 251) And when former students have become teachers, they pass the torch: “Professors remember the survey and recommend it to PhD students. PhD students or lecturers recommend them to students somehow. And just like that, the survey keeps a reputation.” (In04, 854) Sometimes, students turn to their professors for advice, when looking for a suiting dataset: “A student might ask: What data would work? And the teachers think, like, yes, oh, well, this survey programme [survey name] has all sorts of things in it. […] And then they say: Maybe you should check this survey [survey name].” (In04, 289) The students who have been introduced to a certain survey programme like this may come to think at a later point: “I did my research assignment with this survey [survey name], now I would like to do my thesis with this [survey name] data, too.”(In04, 275) Later on, this interviewee added: “But it is indeed like, it’s like, many people have this reflex: Let’s look at [survey name] first!” (In04, 302)

Preference of a certain survey also seems to be tied to specific topics, in the sense that particular surveys are known for their coverage of topic A or B: “We do notice that there are datasets for any specific topic that are requested over and over again. Or they are used, cited over and over. And then there are other datasets that are maybe not that bad either, on similar topics, but they are barely requested. […] Because they are less popular.” (In01, 536) What is more, there are these surveys that have a monopoly on certain topics: “Yes, well, there are of course survey programmes that occupy a topic almost exclusively. I am thinking about [survey] from [research institute], for example. There is nothing comparable

to that. Here or there may be a single survey on [topic A] or [topic B] or the like. But this specific survey, which is also very good methodologically, from [principal investigator], there is nothing that compares to it." (In01, 593) However, with growing experience and analytic skill, users seem to open their focus for other surveys, in particular, if they want to compare their data with data from other surveys: "And it is indeed in many cases that questions have been asked in three or four data collections at the same time. This is not without reason. They [the users] should be able to compare their results." (In05, 238)

COMMUNITY involvement by BACKGROUND and SKILLS

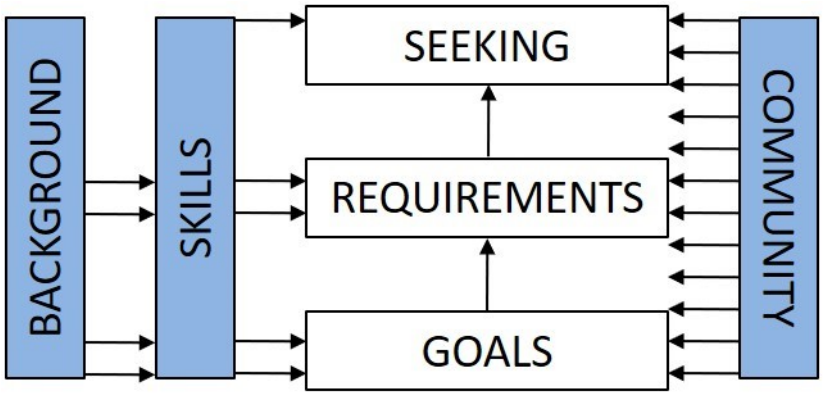


Figure 13 Background, skills, and community

The more experienced survey data users In particular do not only benefit from the community driven data production. They contribute to the community through quality improvement of data and documentation. As one interviewee explained: "You see, not only do we assist our users, they help us, too. Take the issue of 'not finding'. If I have a message for someone, I am the one who is responsible to get that message across. We always look at these issues thinking: why not? Meaning, to improve ourselves, our documentation, our products, we need [that], this how the users react to us." (In05, 802) The interviewee went further in explaining that not only secondary users' requests were important: "You know, as a matter of priority, we are responsive to our users' ideas. How to do documentation, what to include in documentation, how to do it in a better way? And, for that matter, you ought to treat principal investigators as users, too. […] Then, the best ideas that we have

implemented, from my point of view – variable reports, the whole range, online tools – have been requirements or ideas coming from principal investigators as users. […] So we take that in. Users are critical, have a request or are utterly bewildered. In order to check ourselves, if we didn't have this, that would be bad." (In05, 812)

To be able to contribute to data quality improvements in the indicated ways, researchers must have reached a certain stage in their research process, where they really start analysing data. As one participant put it: "We have these requests concerning errors [in the dataset]. [...] And then we go and check the data and the documentation to see if what the user says is actually true. [...] And then, quite frequently, there are errors to be found. [...] And these are things that users find quite frequently when they embark on data analysis." (In04, 83) Usually, users who find and report real errors in datasets are not only further along in the research process, but also more advanced or experienced researchers in general: "Users begin to find such things when they really embark on analysis, that is to say, when they are dealing with variables that are of high importance for their research problem. They find these things, because they have complete understanding. They are expert users. [...] they look at every variable, at correlations between all variables, and only then they start with their final data analysis. And this is where inconsistencies are to be found." (In04, 590) One interviewee even indicated that there were researchers in actual pursuit of inconsistencies in datasets: "There are these large scale harmonisation projects that started to search our data for errors. This development is gaining ground; it is becoming some kind of sport." (In05, 121)

While outsiders may even lack any understanding of what a survey dataset looks like, expert users know their dataset of choice to a point where they are extremely versant with it. What is more, from the interviews it became clear that experience is a major factor in how involved users are in a dataset community. Experts act as true members of a dataset community and develop personal relationships with other community members. Members of a dataset community engage in frequent exchange about the dataset, seemingly in an effort to contribute to its usefulness or even quality. This even seems to be true with regard to the said harmonisation projects: "There was this harmonisation project, and they published, they have their results: what have we found in [these surveys] with regard to data quality? That is a publication – we had to react on it. That much was clear: this is of value for

all researchers, meaning, the researchers have started it, they said, we do quality checks on the large data collections [...]. It all really came on the initiative of the research community." (In05, 203)

How responsible and confident data users may feel when they are reporting errors is reflected in a user's message that one of the interviewees quoted in the context of errors in datasets: "Usually it is like, we get a rather defiant e-mail. The last time it came from [name], a professor at [university]: Dear [survey] team, you screwed things up again – colon – well, and then, wrong labels on the filters." (In04, 603) Sometimes, if they find more complex errors, users provide a small syntax that they have created to correct it. A participant assumed that these users were happy to share their work to the benefit of others: "Because, they [the principal investigators] have a primary interest in the widespread use of their data. […] You know, they are users of a certain kind." (In05, 819) Alongside these dedicated principal investigators there are other highly experienced users who contribute to data improvement. In a sense, they become co-creators of datasets. They are expert users who know all the variables and all the flaws in a dataset.

Then again, while the more experienced researchers find errors and contribute to data quality, other users falsely report errors or other problems with data. For example: "They find errors that aren't there. […] They have something, came to some kind of results in some way that can't be correct. But I often can't even re-enact how they came up with these results. Usually you don't find out what has happened there, unfortunately. Did they just use the wrong dataset, somehow? Or whatever. So, weird things are indeed happening." (In03, 450)

It is not surprising that users make mistakes when they are working with data. It is common, however, that datasets indeed contain errors, for example missing variables. At the very outset of a data use case, the user cannot know, whether there are any errors in the dataset. This means that encountering errors is always a possibility. However, making mistakes when working with data is always a possibility, too. So, there is a point in certain usage scenarios, where a user encounters an inconsistency of some kind. Possible reasons are either an error in the data or a mistake on the side of the user. Presumably the user will at first check their own procedures for possible mistakes. If they find a mistake, they will correct their path and

continue work with the data. If they do not find a mistake, perhaps having taken multiple loops or considered multiple sources of mistakes, they will assume an error in the dataset. What is happening here can be described as activities of *verifying* as identified and defined by Ellis et al. (1993). In an effort to make sure that they are working with accurate data, users get involved with others (e.g., data service staff) to verify: "'See, all this isn't working. My model produces wrong signs systematically. I don't understand this. All of this must be wrong.' And then we look at it, check it and say: No, everything is correct. The signs in your model might be there for a substantive reason. You have to look at it again." (In04, 609)

At least in some of these cases users miss rather obvious mistakes that they have made and quickly assume errors in data instead and resort to data service. As one participant put it: "And then they have requests like: 'But where do I find this question that informs me about this age-set?' And then I start by just looking into the raw dataset or in the questionnaire and find it within just one minute. Then I ask myself, how well or bad has someone done their research? That happens, too." (In02, 532) One interviewee assumed that "there are users who […] are not patient enough to scroll through the documentation" (In05, 100), while another one said: "Well there are just these users who have this understanding: just a quick e-mail." (In04, 498) In comparison, one participant noted: "Maybe, we don't have everything explained a hundred percent clear [in the documentation]." (In03, 490) The interviewed intermediaries were well aware that sometimes, this kind of requests resulted from deficient documentation: "In many cases, the problem is ambiguity of the variable documentation." (In04, 96)

Problems with documentation seem to be an important aspect of people's attempts to find or to work with data. As mentioned above, datasets from large survey programmes, which are explicitly produced for secondary use, usually are very well curated and come with extensive documentation. They are the most findable and understandable datasets in the field of empirical social research. But it seems that even if high quality documentation is sufficient for experienced users, it is less helpful for beginners (cf. J. Niu and Hedstrom 2009). As one of the interviewed curators of a large survey programme indicated, some of the less experienced users still have problems with finding and understanding data on their own, in spite of extensive documentation: "And even though they might find the dataset, they don't notice the tab that says 'documents' right next to it. Or maybe they believe that

perhaps the codebook isn't all that important for them. [...] Maybe just because no one has taught it them yet." (In06, 156) The same interviewee suggested: "From time to time it would be helpful, if younger users especially would just take a look at the documents provided." (In06, 119)

However, other interviewees suggest that the problem is not so much that people are unwilling or impatient; maybe, the extensive documentation is just too overwhelming for them to see through and extract the relevant information: "That is to say that, actually, we try to publish all information that we have and that we deem important to publish and to not keep inside. But, from our point of view, this makes everything absolutely complex, really. And it is very, very difficult for many users to find this information, just like that. Because they cannot see through." (In05, 102) The interviewee added an example: "They reach into the [data catalogue] and find an outdated survey, where there is this note that says: 'This dataset is not available anymore, please go to the recent surveys.' All these are very difficult processes that they cannot understand." (In05, 109) As another interviewee put it: "And yes, sometimes it may be just like, that they are just [overwhelmed] by the abundance of information and access points, meaning that they don't know exactly what is relevant in their special case." (In06, 169) Meanwhile, data curators try to respond to this problem by considering carefully how much and which documentation to publish: "Because then, it is all about pondering: on the one hand, we had this aspiration of answering all questions we could think of through web information in an *ex ante* way. But, as a result, the web information or websites are plenty. And so, at the same time, I can say, yes, well somehow you could have read here, and here, and there, and combine things." (In03, 199) With regard to supposed unnecessary requests and considering this situation of information overload, the interviewee came to the conclusion: "There really are requests that are justified." (In03, 203) The same interviewee said: "And then it is indeed very important that they have an easy way of contacting us, even with stupid questions, because the consequences [of not doing so] could be grave." (In03, 490)

Apparently, problems in finding the right data are not solvable by just offering extensive documentation. By extension this also means that interpersonal exchange of information plays a particularly important role in data seeking behaviour and that the existing technical infrastructure is not sufficient to help with all respective information needs.

Inexperienced researchers are not the only ones to benefit from personal contacts in information seeking. In fact, knowing the right people seems to be especially helpful for experienced researchers who have particular requirements such as datasets with spatial references or in other ways limited anonymization. Doing multi-level analysis with geo-spatial references is currently very popular in empirical social research: “Over the last year or maybe one and a half or two years, it has become a downright trend to do this linking. They all want to link [data] somehow and preferably [...] with small regional units.” (In03, 641) This kind of research is done by advanced or expert users: “Well, concerning analysis of georeferenced data, we usually deal with very well-versed users. At the student level, they cannot do this, as a rule. Because, if you want to estimate a multi-level analysis, you just need considerable know-how in methods of empirical social research.” (In04, 190) Suitable data for this kind of analysis are not plenty. Looking for this kind of data often prompts even expert users to make data service requests like this one: “I am planning this project, using georeferenced data as well. There isn’t enough in [survey], and there is nothing at all in [survey]. Do you know a dataset that also covers this?” (In04, 327) Finding suitable data to solve these specific problems can be very difficult, even for experts.

Another problem with these data requirements is that access to these datasets usually is restricted or not provided at all to protect data privacy. It seems, however, that if you know the right people, there may be ways to work with these data after all: “What happens is that people read the paper from [researcher] and call us: ‘How did he do that? Where are these data? I can’t find them [in your catalogue].’ And then I have to say, ‘sorry, [researcher] has a direct line to [field institute]. I can’t give you the data, we don’t have them.” (In04, 391) Apart from legal restrictions, there may be other cases, where researchers require data services that are not legally problematic but entail extra effort on the part of the data provider. Data providers may be quite inclined to attend to these requirements if they can expect support or cooperation in the future: “Well, these reciprocities, they make, they are, to my mind, very important in the data business” (In04, 369), one participant indicated.

As for data with access restriction, it became clear at one point that researchers commonly share them among each other instead of officially requesting access: “This is mainly true for datasets that are not freely available. They are shared on a personal level, from one person to the next. But this is really common.” (In04, 403) Informal sharing of data between

researchers or students is an easy way to avoid requesting official access, but it leads to a whole range of problems. Apart from privacy or copyright issues there are problems that result from sharing different versions of datasets, including outdated or corrupt files. These may result in wrong analyses, problems with comparability and with replication. The interviewee gave an example: "[survey] could not be published correctly. This is why there are five different versions of the dataset out there. And you can see this in the publications, because they have different sample sizes." (In04, 453) But some users seem to have a casual way of dealing with these problems: "They have this impetus of 'first, I take a look at it and when I see that there is something, only then will I request access to make it official.' When I was a student, at [university department], people had loads of illeg[itimate] data […]. Data that they were not supposed to have. And that is just the way it is: first, you take a look at the data. And then you say, no one says, if they are asked: 'I have this or that dataset, is anyone interested?' No one says 'no' to that, but they make a copy, just for now." (In04, 408) The interviewee added that, even if there is no access restriction and the data is downloadable for free, users tend to share these datasets informally, for example in study groups. Apparently, users just go with what is the easiest way for them.

3.2 Findings

This subchapter presents the key findings that have been drawn from the results of the qualitative study (subchapter 3.2.1). It describes the developed grounded theory by drawing on the analyses of the key codes and categories. This account leads to five hypotheses that form the basis for the quantitative study. To give recognition to those codes and categories that had less relevance for the developed theory, this chapter ends with an account of a few other findings that will not be part of further analyses but are worth revisiting in later studies (subchapter 3.2.2).

3.2.1 Key Findings and Hypotheses

The research question defined in the first chapter was: What are the characteristics of social scientists' information seeking with regard to survey data? Following Wilson (2002), information seeking behaviour was defined as *goal-oriented problem solving* and was further specified as

- depending on individual characteristics and context
- occurring in stages, cycles or patterns and

- encountering barriers.

To find out how this phenomenon is shaped with regard to survey data, a qualitative study has been conducted, based on seven areas of exploration:

- The users' educational/professional background
- The users' research experience and data literacy
- Goals, needs and purposes of users
- Requirements (data quality, topics, methods) when looking for data
- The role of documentation
- Information sources and channels (the role of intermediaries and information technology)
- Barriers and problems when looking for data

Conducting and coding the interviews with grounded theory methods yielded rich and multifaceted data in all seven areas of exploration and with regard to the research question. This research resulted in a grounded theory of information seeking behaviour of survey data users that can be described as "Theory of problem-solving by community involvement". The key findings given in the following paragraphs are based on this theory.

Figure 14, "Model of problem-solving by community involvement", provides a visual representation of this theory (a larger version of the figure is provided in Annex 16).

The visualisation provides detailed information on the relationship and correlation between the core categories of the theory. It is a derivative of the schematic representation of the core categories given in Figure 7. The Model implies that the core category, COMMUNITY, is influencing all other categories. The spectrum of community involvement is depicted by the colour gradient that increases in depth towards the right end of the spectrum ("high community involvement"). All the other categories vary in their manifestation according to the degree of community involvement. The most linear correlation is visible between COMMUNITY and GOALS, which is depicted at the bottom of the model. Ambitious GOALS correlate with high COMMUNITY involvement. The BACKGROUND and SKILLS categories are depicted in the centre of the model and are labelled with the indicators "experience, seniority, data literacy, skill". On the far left of the spectrum are people with no experience

and no survey data literacy. On the far right of the spectrum, there are very experienced, skilled people, who have high goals and profound community involvement.

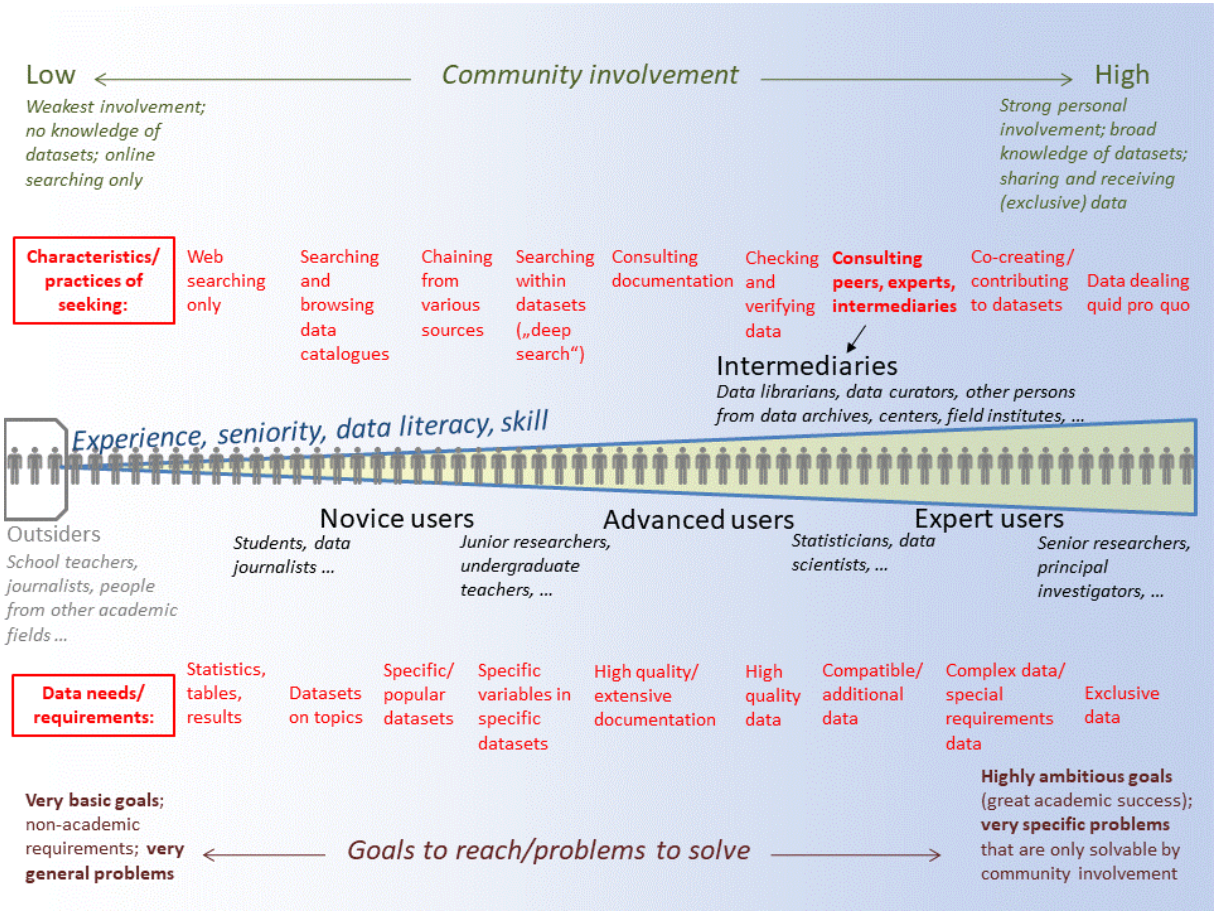


Figure 14 Model of problem-solving by community involvement

The categories SEEKING and REQUIREMENTS are each represented horizontally along these spectra. Characteristics and practices of SEEKING as well as data needs or REQUIREMENTS are listed in an order that represents their probable occurrence with regard to the degree of community involvement, experience, and goals. The interviews revealed that goals and problems of secondary data users are very diverse. Some of the goals that shine through in users' requests to data service are: getting published; achieving academic success; learning how to work with data; graduating. Some of the problems are: lack of suitable data; lack of recent data; lack of information on data; lack of skill; problems with data quality. All these and other problems trigger people to seek information that helps them to resolve their *problematic situation*, understood as discrepancy between their life-world and encountered phenomena (Wilson 1999). The interviews revealed several insights on how the data users

proceed to seek that information and what characteristics and which other factors constitute their seeking behaviour.

The most interesting finding in that respect turned out to be the relevance of the category *community involvement*. Community involvement seems to play a significant part in researchers' data seeking behaviour, in particular with regard to goal-orientation and problem solving. The interview data suggest that **personal interaction with others is a significant factor in goal development, goal achievement and problem resolution for researchers who want to reuse survey data**.

Personal interaction with others while looking for data is to be contrasted to information seeking patterns that merely involve individual activities such as searching the web, searching databases, and searching complementary material such as dataset documentation. While individual activities like these are predominant or even sufficient in literature seeking, informal information activities through personal interaction seem to be imperative information practices in data seeking. With regard to survey data, **informal information seeking by personal interaction is facilitated through the existence of vital communities surrounding large survey programmes**. These communities are each made up by people who fulfil different roles with regard to the survey; the community comprises the survey's principal investigators and other primary researchers, people in the field institutes (interviewers, coordinators etc.), data managers, data curators, data librarians, and the data users. Within a community, the survey's datasets as well as complementing information on these datasets (documentation) are produced, shared and used. Some community members play different roles at once, for example, they are creators/co-creators and users of a dataset. Sometimes secondary users apply themselves in data improvement, for example, when they detect and report errors in the data. And sometimes, principal investigators make suggestions for improvement to data curators on behalf of secondary users. The same person can have different roles in different survey communities.

The most visible and productive communities are those surrounding the large survey programmes. Since these programmes produce the most frequently used datasets, further investigation is focused on these communities.

Dataset communities emerge and persist, because knowledge of them is handed down from senior researchers to junior researchers or shared between peers. Typically, students are introduced to one or more large survey programmes during their education in empirical social research. They tend to revisit these surveys, when they are looking for data that they can use in their research assignments. Their supervisors also encourage them to use these data. In this way, young researchers gain knowledge of data infrastructure services such as data repositories or data archives. From there, they find access to even more data from other surveys. More advanced researchers might also find new or other datasets through interaction with peers, for example, at conferences. However, some researchers repeatedly work with data from one survey and do not look for alternatives throughout their careers. They are actively searching for new dataset versions or complementary datasets from their survey of choice. They are not interested in working with data from other surveys. For these users and other interested researchers, large survey programmes offer services such as exclusive conferences or "meet the data" workshops for users of their data.

Overall, it seems that **with growing experience, seniority, and data literacy, community involvement is increasing**. The broad spectrum of backgrounds of people looking for data starts with outsiders who have no experience with survey data analysis, such as school teachers or researchers from other disciplines. It goes on to novice users such as students or graduates (among them data journalists) with knowledge and basic experience in survey data analysis and further to advanced users whit solid data literacy and experience. On the far end of the spectrum there are expert users who have rich experience and knowledge in data analysis. Expert users are often primary investigators or contributors to survey data. While outsiders show no community involvement at all, novice users already had first contact with data providers and other data users, advanced users maintain and advance their networks, and expert users find themselves at the core of survey communities or even initiate them. While outsiders have no knowledge of datasets and how or where to find them, experts have a broad knowledge of datasets and are actively sharing and receiving data, sometimes even exclusively. The outsiders' community involvement is very weak and completely impersonal, while expert users show a strong personal community involvement.

The users on the spectrum from outsiders to expert users have different goals to reach and problems to solve. The outsiders' goals are very basic, their requirements non-academic,

and their problems are very general. And while novice users like students have rather moderate goals (e.g., graduating) associated with problems of little complexity, expert users have highly ambitious goals (such as great academic success) which present them with very specific problems. Here are some generated examples of goals and problems that users on the spectrum from outsiders to experts may have:

- **Outsider's goal**: obtain empiric results on topic X; **problem**: lacking skill to read and analyse datasets.
- **Novice's goal**: doing a research assignment on topic X; **problem**: apparent lack of suitable data.
- **Advanced user's goal**: getting published; **problem**: data quality (e.g. not enough cases in a sample)
- **Expert's goal**: innovative and outstanding findings; **problem**: legal or ethical barriers.

Expert users need to make complex analyses with complex data and they need to make analyses that no one else is doing or has done before. This is also reflected in the data that they need or the requirements that the data should meet; while outsiders often do not want to deal with survey datasets and need mere statistics or results, more experienced users are looking for datasets that fit their subject interest or even more likely, datasets of specific or popular surveys. The more experienced users are, the more specific their problems and hence, their requirements. For instance, advanced users search for specific variables in known datasets, sometimes with the intention to pool these data with other data. To do so, they need and appreciate high quality and extensive documentation. At this point, data quality becomes an important requirement as well, because the more specific the research question, the more difficult it will be to find data with appropriate sampling (for example, if your sample population is "working single mothers in urban environments who vote for a specific party"). Finally, expert users need datasets with special features such as spatial information. The spectrum of needs or requirements can be illustrated by these generated example requests:

- Outsider: "I need empirical findings on topic X."
- Novice user: "I need a dataset to answer my thesis research question."
- Advanced user: "I need the most recent wave of this dataset."

- Advanced user: “I need a cumulative dataset of all waves.”
- Advanced/expert user: “I found an error in the dataset and I need it to be fixed.”
- Expert user: “I need this dataset with other area codes.”
- Expert user: “I want to do spatial analyses as Dr Y has done them, but the dataset provided doesn’t contain the geo-references.”

Searching for datasets on topics of interest (in the sense of a conceptual keyword search) seems to be an activity that is more prevalent among novice users than among experienced or expert users. One reason may be that more experienced researchers already know which surveys cover certain topics. Another reason may be that conceptual searches for datasets are often unsuccessful, because there is no standardized conceptual indexing of survey datasets available. These findings are in line with recent research from Guangyuan Sun and Christopher S.G. Khoo who found that more experienced researchers turn to familiar data archives and tend to search known datasets, while the less experienced carry out subject searches or browse data catalogues (Sun and Khoo 2017). Browsing in particular seems to be a preferred strategy to "gain an overview idea of what kinds of datasets are available" (Sun and Khoo 2017, 69). Sun and Khoo also point to issues with knowledge representation in data documentation (Sun and Khoo 2017). Apparently, more experienced researchers have learned that good knowledge of particular surveys is more helpful than conceptual searching when looking for data to analyse topics of interest. In fact, as interviewees have indicated, researchers tend to develop their topic of interest with an already known survey at hand.

How users are looking for data in terms of characteristics or practices of seeking is equally depending on their experience or seniority. People who lack experience with data analysis and are primarily interested in statistics or results perform *web searching* only. When they have acquired some data literacy and understanding of survey data, they may even *search* and *browse* data repositories or data catalogues. *Chaining* from sources like mass media outlets is also employed by novice users. *Chaining* from journal articles or even *searching* within known datasets (“deep search”) seems to prevail among more advanced users. For these users that are interested in specific variables, *consulting the documentation* provided with the dataset is necessary. They also perform activities such as *checking* and *verifying* data, as soon as they have found relevant data. The most experienced users are actively *contributing* to datasets if they find inconsistencies. Expert users *co-create* the data that they

want to use by either helping to improve them (with added value documentation) or even by being principal investigators. In certain circles of a community, users seem to find and get access to exclusive data, sometimes in *exchange* for other data or other forms of cooperation. An information seeking practice that seems to be employed by all users is *consulting* peers, experts, and intermediaries such as data librarians. Again, this practice of employing personal contacts to find data seems to occur more with advanced or more experienced researchers than with outsiders or novices. A recent study by Sheila Pontis and colleagues that investigated the scholarly activity of keeping up to date, albeit not with a focus on research datasets, yielded similar results regarding the connection between experience and reliance on personal information seeking practices (Pontis et al. 2017). Similar results have been reported from another qualitative study of social scientists' data reuse behaviour by Ayoung Yoon, who concluded that using scholarly lineage and networks to learn about and gain access to data is common practice, especially for more advanced researchers (Yoon 2017). A reason for this correlation may be that problems of outsiders or novices could more easily be solved by just looking at the documentation or information provided online.

It seems that, with more ambitious goals and more specific problems community involvement becomes more and more necessary, because **community involvement facilitates goal-oriented problem solving with regard to survey data**. In a dataset community, there may always be someone who knows even more about specific issues with the dataset than anyone else. In a well-established dataset community, the community members know who to contact for advice. For example, community members will refer to other members who they know and who they interact with. Basically, problem solving works through the social networks of dataset communities. Intermediaries (e.g., data librarians) or other central figures have particular network knowledge (that is, knowledge about the community).

Being an active community member can improve a researcher's outcomes. Working outside the community can be successful as long as the problems at hand can be solved by consulting the basic information provided alongside the dataset. Usually there is plenty of information or dataset documentation available online. But as the interviewees have indicated, not all possibly relevant information can be included in the documentation.

Certain people may have information that is crucial to work with a dataset or they may be the only ones who know how to find this information. So, people with rather challenging problems will find that they need informal ways of problem solving or information seeking. For instance, if they need datasets that contain sensitive information, another community member might be able to tell them where to find these data, or who to approach with this issue. Information like this has a confidential character and is usually not available online. It seems that, for the more general problems, documentation that is available online might be sufficient in order to find and work with data. **The more specific and delicate the problems are, the less likely it seems to be that they will be solvable by merely relying on the information provided online.** What is more, with growing data volume and data complexity (comparative survey programmes, panel surveys, etc.), data documentation cannot be comprehensive in every way. Not all potential use cases can be anticipated by the data managers or data curators. As the interviewees have indicated, some documentation is only available because other community members have made a contribution. Also, datasets tend to have flaws or errors that can only be resolved by getting in touch with others, ideally with data managers, data curators, or principal investigators. Only by contacting these people, researchers who have found inconsistencies can make sure that they work with an authoritative, verified dataset. In that regard, community involvement and shared responsibilities in a community can have positive effects on data quality, which benefits all community members. With regard to personal benefits, it seems that those who maintain good personal relationships and informal contacts to other community members are more successful in getting the data that they need.

It is noteworthy that, on the far end of this spectrum, there even seems to be the peril of community involvement shifting towards **data dealing within exclusive circles**. Information sharing behaviour of this kind can be unfair to less involved researchers. It may even hinder transparency and foster biases in research. But even outside these highly exclusive circles, **informal sharing of datasets among peers** can be perilous: if for certain studies different versions of datasets are circulating, researchers are working with outdated or otherwise not authoritative versions of datasets. This situation of different people working with different datasets complicates comparison of research results and compromises scientific accuracy.

In sum, the qualitative study on information seeking behaviour of survey data users led to a range of interacting assumptions regarding the grounded theory of problem-solving by community involvement. From these assumptions, hypotheses for the quantitative part of the investigation were drawn (Table 3).

Table 3 Hypotheses on data seeking practices and community involvement

(1) The data seeking hypotheses:
(1a) When looking for data, information seeking through personal contact is used more often than impersonal ways of information seeking.
(1b) Ways of information seeking (personal or impersonal) differ with experience.
(2) The experience hypotheses:
(2a) Experience is positively correlated with having ambitious goals.
(2b) Experience is positively correlated with having more advanced requirements for data.
(2c) Experience is positively correlated with having more specific problems with data.
(3) The community involvement hypothesis:
Experience is positively correlated with community involvement.
(4) The problem solving hypothesis:
Community involvement is positively correlated with problem solving strategies that require personal interactions.

The hypotheses reflect the following findings from the qualitative study: When looking for reusable survey data, information seeking practices that involve personal contact are very important (Hypothesis 1a). The choice of personal and impersonal ways of information seeking seems to be related to seniority or experience in survey data research (Hypothesis 1b). The core concepts of goals, requirements, and problems that constitute information seeking practices are related to experience as well: More experience is associated with more ambitious goals, advanced requirements, and more specific problems (Hypotheses 2a, 2b, 2c). At the same time, more experienced researchers are more involved in survey data communities (Hypothesis 3). This community involvement enables researchers to solve their more specific problems by means of problem solving strategies that require personal interactions (Hypothesis 4).

3.2.2 Other Findings

The data collected in the qualitative interviews yielded even more insights to secondary users' data seeking behaviour. These aspects are worth reconsidering in future studies and some of them could be used as additions to the presented model on "Goal attainment and problem solving by community involvement". The most relevant of these findings are briefly reported here.

The feeling of commonality in a dataset community. Community members contribute to dataset improvement by sharing their problems or even their solutions with other community members. There may be different reasons for that, for example, a sense of commonality between community members based on their work with the same datasets (similar to research networks).

Shared responsibilities in dataset communities. It is rarely the case that large surveys are conducted, curated, and distributed by the same people or institutions. Usually, there are shared responsibilities in the creation of a large survey programme. In particular, the cooperation between data curators and principal investigators can be very close, they may even share work tasks to provide datasets. On the one hand, shared responsibilities and community involvement can have positive effects on data quality, for instance through double checks. On the other hand, communities with shared responsibilities can also be challenging or confusing to users, because the whole community is never visible to the user. Knowledge about roles and responsibilities within a dataset community depends on experience, seniority or involvement in the research community.

Users' tasks, requirements, and goals. The interviews have rendered many examples of tasks, requirements, and goals that users have revealed to the interviewees. For example, interviewees have indicated that some users do not need datasets but are only interested in the survey questionnaire, because they are planning to reuse questions in their own survey. Another example is that some users are looking for data on trending topics (e.g., 'migration'). Another interesting finding is that there seem to be researchers who are interested in working with unique datasets, which somewhat contradicts to the finding that users tend to use popular surveys. Finally, it seems that there are very few researchers who are interested in just doing exact replications of analyses that others have done.

Barriers in data reuse. Interviewees have indicated several types of barriers encountered by secondary users of survey data. In general, many barriers relate to lack of experience or skill in data analysis; these cases are included in the theory and model presented above. Another interesting finding with regard to barriers in data seeking is that users who are looking for data are overchallenged by the diverse research data landscape. This includes not being able to tell what data exist, but also being challenged by the lack of standardisation in documentation. On the grounds of the theory presented above, it seems that these problems are evaded by sticking to known datasets and involving oneself with the community. It would be interesting to investigate though, in what way better findability of data could affect data seeking behaviour. Finally, for some users there are also technical barriers to data reuse, such as lacking facilities or software.

3.3 Validity of the Results: Respondent Validation

Respondent validation, a common procedure in qualitative as well as mixed methods research (Bryman 2012; Torrance 2012), was used to assess the validity of the developed theory. Respondent validation was carried out by presenting the diagrammatic model (Figure 14 or Annex 16) to two participants of the study and asking them to assess its accuracy.

3.3.1 Respondent Validation: Sampling and Design

The two participants who were included in the respondent validation were interviewees no. 4 and no. 5. This sample was based on the fact that these were the interviewees who had made essential contributions with regard to the concept of community involvement, which is central to the theory that has been developed.

Both respondents were interviewed separately. They were invited to one-hour meetings that took place on October 11 and October 12 2018. The time frame of one hour turned out to be sufficient. The respondents were provided with a copy of the diagrammatic model (Figure 14 or Annex 16) two days before the meeting. This copy was sent to them by e-mail. The e-mail contained the following additional information: “In preparation of our meeting on [date], I am sending you a diagram that depicts the theory of “problem-solving by community involvement” that I have developed in my qualitative study. The theory describes essential factors that are supposed to affect information seeking behaviour when looking for data.” This e-mail contained no further explanatory remarks on the diagram.

To guide the conversation, a short introduction to the theory, the qualitative study, and the diagram was given in the beginning of each of the two sessions. A written version of this introduction is available in Annex 17.

3.3.2 Respondent Validation: Results

The first respondent (interviewee no. 5) confirmed the idea that community involvement is an essential factor of information seeking behaviour with regard to survey data. She found it plausible that people with higher expertise and seniority in survey data analysis are more involved in this community, and that the community involvement is helpful when looking for data. Two points of criticism stood out in this interview:

(1) The respondent did not entirely agree with the assignment of specific user groups along the spectrum of experience, seniority, data literacy and skills. In particular, she found that the placement of journalists and students might as well be further right down the spectrum, meaning that there are very experienced and very involved journalists or students to be found.
(2) She completed this view by expressing irritation over the rather judgmental categories of “no knowledge of datasets” or “very basic goals” as opposed to “broad knowledge of datasets” or “highly ambitious goals”. The respondent asserted that these judgements seemed to refer to very general aspects, such as skills or goals. She explained that, for instance, journalists or researchers from other fields have ambitious goals as well, albeit not in the area of survey research.

The first point of criticism (1) was put to test in the quantitative study. It refers to hypothesis 2a that was tested by correlating measurements of experience with measurements of goals (see subchapter D.6.2.1). The second point of criticism (2) revealed a weak point of the diagram and the explanation given with it. Of course, the goals, skills, etc. specifically refer to the knowledge and use of survey datasets. Lacking knowledge and skills or basic goals in this area do not indicate general low qualification or lack of ambition. For the quantitative study, the operationalisation of skills was clearly directed towards methodological and analytical competences in survey data research (survey data literacy). The misunderstanding was taken into account for the interview with the second respondent, who was presented with a more specific explanation of what experience, skills and goals in the context of this study meant. The irritation did not reoccur with the second respondent.

The second respondent (interviewee no. 4) also confirmed, that community involvement was an essential factor in survey data seeking. With regard to the presented diagram, the respondent pointed out that the general framework of correlating community involvement, experience, and goals/problems was very convincing. He also found that the data needs and requirements were covered very accurately. The characteristics and practices of seeking depicted in the diagram were also accurate, but not necessarily in the depicted order. Maybe, they could not even be viewed as clear and separate categories, but should rather be perceived as intertwining. In particular, the practice of consulting peers, experts, and intermediaries seemed difficult to correlate with stages of experience or community involvement. The respondent suggested that the characteristics and practices of seeking should be relatable to all stages of experience, to all grades of community involvement, and to all levels of goals or problems. The practice of consulting intermediaries in particular should occupy a special position, because the work of intermediaries (such as data curators) is also fundamental to various practices of seeking such as consulting documentation, searching and browsing data catalogues, etc.

This criticism was also put to test in the quantitative study. It mainly refers to Hypothesis 1a that was measured by univariate analysis of practices of information seeking. The quantitative analyses did not support the assumption that data professionals (data archive staff, data librarians, data curators) have a special role in this regard (see subchapter D.6.1.1). This can only be confirmed for intermediaries in a wider sense, which would include professors, supervisors, colleagues, and friends.

In conclusion, the results of the respondent validation support the main aspects of the grounded theory of problem solving by community involvement. Details of the diagrammatic model were questioned to an extent that was manageable within the tests that were already planned for the quantitative study.

4. Summary

The qualitative part of this study has successfully produced a grounded theory of information seeking behaviour of survey data users. For this study, six experts in data service were interviewed on their experience with survey data users' information seeking

behaviours. The experts were sampled by means of initial and theoretical sampling (see subchapter C.2.3). The interviews followed an interview guide covering eight interview topics that were based on seven areas of exploration that had been deduced from preliminary theoretical considerations (see subchapter C.2.1). All interviews were recorded and transcribed. The transcripts were coded and analysed using the constant comparative method (see subchapter C.2.4). The processes of open and focused coding were accompanied by memo-writing that was oriented towards theory-building (see subchapter C.2.4.2). Scrutinizing and analysing the data in this way, six categories have emerged as promising cornerstones of the grounded theory of information seeking behaviour of survey data users. These are the categories of background, skills, goals, requirements, seeking, and community (see subchapter C.3.1). In the theory, categories build on other categories, and some categories impact others in specific ways. The nexuses between the categories are briefly summarized here:

Users come from different backgrounds and have different skill sets with regard to data literacy. How users are looking for data in terms of characteristics or practices of seeking is depending on their experience and data literacy. Users on a spectrum from outsiders to expert users have different goals to reach and problems to solve. Their requirements when looking for data differ accordingly. Personal interaction with others is a significant factor in goal development, goal achievement, and problem resolution for people who want to reuse survey data. Information seeking by personal interaction is facilitated through the existence of data communities. These communities emerge and persist, because knowledge of them is handed down form senior researchers to junior researchers or shared between peers. With growing experience and data literacy, a users' community involvement is increasing. Community involvement facilitates problem solving with regard to survey data and thus, being an active community member can improve a researcher's outcomes. The more specific and delicate the problems are, the less likely it seems to be that they will be solvable without community involvement.

Based on the core findings of the analyses, the developed grounded theory was named "theory of problem-solving by community involvement". A diagrammatic representation of the theory is provided in Figure 14. To exemplify the qualitative findings with survey data collected in the quantitative study (chapter D.), hypotheses as provided in Table 3 were

drawn from the grounded theory. The survey design and questionnaire for the quantitative study were developed on the grounds of these hypotheses.

D. Quantitative Study

1. Methodology and Research Design

The quantitative study was designed as a web survey of survey data users. The aim of this survey was to gather data for exemplification and testing of the grounded theory that had been developed in the qualitative study.

The theory makes assumptions about secondary survey data users' information seeking behaviour. Obviously, secondary survey data users form a population that is impossible to identify in its entirety. Therefore, a proxy population was surveyed here to exemplify the theory. This survey population is made up by actual secondary users of survey data, which is assured by their registration with an online catalogue that provides access to survey datasets for secondary analysis. It is assumed here that people who go through the trouble of registering with the data catalogue (including a confirmation step) are in fact interested in reusing survey data. With this approach it is neither possible nor intended to calculate inferences to the general population of secondary survey data users. All descriptives and inferences presented here are therefore restricted to the survey population. Regardless, the exemplification will still provide insights that may prove useful in other or more general contexts. More details on the survey population are given in subchapter D.0.

The survey questionnaire was developed on the grounds of the theory and the resulting hypotheses (Table 4). The questionnaire was designed along the core concepts of these hypotheses. The core concepts (as highlighted in the hypotheses) are: practices of data seeking; experience; goals; requirements; problems; community involvement; and problem solving. **Community involvement** as the theory's key category serves as an important independent variable for problem solving. Another important category is **experience** which is expected to be associated with having ambitious **goals**, advanced **requirements**, and very specific **problems**. **Community involvement** is expected to grow with experience. It is further expected that community involvement **facilitates problem solving** and thus information seeking with regard to survey data. The core concepts will be defined and explained in depth in the next subchapter.

Table 4 Hypotheses on data seeking practices and community involvement

(1) The data seeking hypotheses:
(1a) When looking for data, information seeking through personal contact is used more often than impersonal ways of information seeking.
(1b) Ways of information seeking (personal or impersonal) differ with experience.
(2) The experience hypotheses:
(2a) Experience is positively correlated with having ambitious goals.
(2b) Experience is positively correlated with having more advanced requirements for data.
(2c) Experience is positively correlated with having more specific problems with data.
(3) The community involvement hypothesis:
Experience is positively correlated with community involvement.
(4) The problem solving hypothesis:
Community involvement is positively correlated with problem solving strategies that require personal interactions.

The questionnaire was created based on the operational definitions of the variables that are addressed in the hypotheses (subchapter D.2). This questionnaire was administered online as detailed in subchapter D.0. Subchapter D.4 contains a description of the sample. Subchapter D.5 describes the index and scale development. Various analyses were made to test the hypotheses (subchapter D.6).

2. Development of the Questionnaire

In order to develop indicators and questions, the key concepts of the theory needed to be defined. This process of operationalisation of concepts entails defining the key concepts and measurable indicators that can stand for the concepts of interest (Bryman 2012).

To prepare the instrument (questionnaire), operational definitions of the concepts (variables) that made up the hypotheses were drawn and dimensions as well as indicators were identified. In completing the process of operationalisation, measurable indicators that can stand for the concepts were derived and transferred into questions for the questionnaire.

The definitions of the key concepts as well as indicators and questions for their measurement are presented in the following paragraphs. The full questionnaire is available in Annex 18 (English) and Annex 19 (German).

2.1 Community Involvement

2.1.1 Operational Definition

For this questionnaire, community involvement was defined as seeking and maintaining formal and informal contacts with other community members in order to contribute to community interests as well as to benefit from the contribution of others.

Survey or dataset communities are made up by people who fulfil different roles with regard to a survey: the survey's principal investigators and other primary researchers; people in the field institutes (interviewers, coordinators etc.); data managers; data curators; data librarians; data users. Within a community, the survey's datasets as well as complementing information on these datasets (documentation) are produced, shared and used. Some community members play different roles at once, for example, they are creators/co-creators and users of a dataset. The same person can have different roles in different survey communities.

Community members contribute to a dataset community in various ways. The most obvious contribution would be to participate in the creation, conduct, and data dissemination of a large survey programme as principal investigator, data manager, or in any other function. A common contribution to data quality is made by secondary users who detect and report errors in the data. Another example are principal investigators who make suggestions for improvement of datasets to data managers or data curators on behalf of secondary users. Common contributions include sharing data publicly (via archives or repositories) as well as privately (e.g. among peers). Some data users share their syntax or improved documentation of datasets, and others give talks or workshops on the use and specifics of particular datasets. A special case of contribution would be to take part in a publicly funded access panel programme. This means to be able to have customized survey questions included in a panel survey under the condition that the resulting data is made publicly available. This case is a mixture of contribution to community interest and benefit from community involvement.

2.1.2 Measurement

Community involvement was measured by the respondents' contribution to community interests. This measurement included a question on data sharing and another question on other possible contributions. The data sharing question was: "Have you ever shared data from your own survey (or from a survey that you have conducted together with others)?" (Q21 Data sharing/if) Other possible contributions were measured by asking: "Some people who are working with survey data contribute to the creation, improvement, or dissemination of survey data for reuse in some way or another. Have you ever engaged in one or more of the following activities?" (Q23/24 Own contribution)

From these community related questions, a scale to measure community involvement was created (see subchapter D.5.2).

2.2 Experience

2.2.1 Operational Definition

The concept of experience traces back to the category BACKGROUND, which is an important category from the qualitative study. For the purpose of the quantitative survey, experience was defined as a capacity made up by actual experience with survey data analysis (data literacy), complexity of applied methods of data analysis (methodological skills), educational background (degree), and knowledge of the survey data landscape. This kind of experience includes aspects of data literacy as well as methodological skills.

Experience is expected to be a major independent variable for the analysis of problem solving by community involvement. With regard to survey data seeking and use, the spectrum of experience ranges from outsiders' experience to expert users' experience.

2.2.2 Measurement

Experience with survey data use was measured with a few different questions. As a filter question, respondents were asked: "Have you ever used survey data for your work or for your studies?" (Q02 Use of data). Respondents who had answered "Yes" were presented with the following question on data literacy: "Have you ever done statistical analyses with data?" Those who had done statistical analyses were further asked: "What methods have you used for data analysis so far?" to determine their proficiency with survey research methods. Regardless of past use of survey data, all participants were asked about their

knowledge of popular survey programmes: “Have you ever heard of the following survey programmes?” (Q07/08 Known data/closed) and “What other survey programmes do you know?” (Q09 Known data/open) The 25 surveys in Q07/08 were comprised by: the most downloaded surveys from the GESIS data catalogue (GESIS - Data Archive for the Social Sciences 2019); the RatSWD output paper on large survey programmes in Germany (RatSWD 2017); additional popular international surveys and key surveys from the United States and United Kingdom; and feedback from the respondent debriefings and pretest (see below). The educational background was measured with the question "What is your highest college or university degree?" (Q28 Degree).

All these indicators were used to create an experience index with four dimensions (see D.0).

2.3 Practices of Data Seeking

2.3.1 Operational Definition

Information seeking behaviour as it is understood here is goal-oriented problem solving (Wilson 2002) that is occurring in *characteristics* or *practices* (Ellis 1989, Meho/Tibbo 2003, and others). Characteristics and practices of seeking that are relevant for the theory of problem-solving by community involvement are depicted in the diagrammatic visualization of the theory (Figure 14). The theory includes personal as well as impersonal ways of seeking. Information seeking without interpersonal contact includes patterns that involve only individual activities such as searching the web, searching databases, and searching complementary material such as dataset documentation. Information seeking through personal contacts refers to mainly personal interactions with other community members. This includes information seeking through known contacts (such as peers) as well as unknown contacts (in particular in online social networks). Sometimes, people who are looking for data through personal contacts are referred from one contact person to another.

2.3.2 Measurement

The knowledge of popular survey programmes (Q07/08) is part of the measurement of experience as stated above. With regard to practices of data seeking, sources of known data were further investigated by asking respondents: “Where do you know these survey programmes from?” (Q10 Sources of known data). Afterwards, respondents were asked about their general data seeking behaviour. As a filter question, respondents were asked: “In the past two years, have you searched for survey data that you could use for your work or

your studies?" (Q11 Seeking data). Respondents who responded with "No" were not presented with the questions on goals, problems, and requirements that are described in the following paragraphs. Data seeking behaviour was surveyed with the question "Which of the following sources do you use to find suitable data?" (Q15 Seeking/sources) Possible answers included impersonal ways of seeking (I search data catalogues) as well as personal ways of seeking (I ask a friend or colleague for suitable data).

2.4 Goals

2.4.1 Operational Definition

In this study goals are defined as the purpose of the users' activity with regard to data. To a great extent, users deal with professional tasks such as completing a research assignment, writing an article, or giving a presentation at a conference. Behind these tasks, there are more general goals such as being a successful student, researcher or journalist. It is expected here that community members on a spectrum from outsiders to expert users have different goals. With regard to survey data, the outsiders' general goals are expected to be very basic. While novice users like students should have rather moderate general goals (e.g., graduating) expert users have highly ambitious goals (such as great academic success).

2.4.2 Measurement

The question designed to measure **goals** with regard to data use was: „What have you needed survey data for in the past two years?" (Q06 Goals/purpose) Possible answers include diverse specific purposes that are not necessarily exclusive. The possible goals or purposes were phrased to represent three levels of ambition as illustrated in Table 5. Obviously, the chosen allocation of the purposes to low, medium or high levels of ambition is not without problems. For example, a policy or strategy paper may well adhere to high scientific standards and contain sophisticated data analysis. Likewise, the use of data for teaching is not necessarily tied to high ambition with regard to survey data analysis. However, both these items can be seen as indicators for academic ambition for which writing of policy papers is less relevant than teaching. The allocation to the different levels of ambition should be read as a rough approximation rather than an exclusive attribution.

Table 5 Goals (purposes) according to levels of ambition

Purpose	Level of ambition
Use of data for scientific publication	High
Use of data to replicate results	High
Use of data for teaching	High
Use of data to come up with research question	Medium
Use of data for thesis	Medium
Use of existing measures	Medium
Use of data for practice	Low
Use of data for policy or strategy paper	Low
Use of data for non-scientific publication	Low

In order to ensure that the current level of ambition of goals or purposes, the question was restricted to a two-year scope.

2.5 Requirements

2.5.1 Operational Definition

For the purpose of this study, requirements are defined as criteria that data should meet to be reusable with regard to the given goal (or purpose). The success of data seeking is determined by whether the found data meet these requirements.

The users' level of experience is also reflected in what data they need or the requirements that the data should meet. Outsiders often don't want to deal with survey datasets and need aggregate statistics or results. More experienced users are looking for datasets that fit their subject interest or datasets of specific or popular surveys. The more experienced users are, the more specific their problems and hence, their requirements. For instance, advanced users search for specific variables in known datasets, sometimes with the intention to pool these data with other data. To do so, they need and appreciate high quality and extensive documentation. At this point, data quality becomes an important requirement as well, because the more specific the research question, the more difficult it will be to find data with appropriate sampling. Finally, expert users need datasets with special features such as spatial information. It is expected here that the requirement to find previously unknown

data on specific topics of interest is more prevalent among novice users than among experienced or expert users, because knowledge of the dataset landscape increases with experience.

2.5.2 Measurement

The question to measure requirements was "When searching these data, how important were each of the following requirements? Please indicate importance on a scale from 1 (not important at all) to 5 (very important)." (Q12/13 Requirements/closed). As with the goals and problems, the list of requirements was developed based on the qualitative interviews and along a spectrum from requirements that outsiders might have (data that are easy to understand) to requirements of experienced researchers (data that hadn't been analysed before).

2.6 Problems

2.6.1 Operational Definition

When looking for data, users are trying to reach their goals, but they are facing a variety of problems. In the present context, problems are defined as intervening events or circumstances that complicate the finding or making use of data. In that sense, problems can also be understood as barriers in information seeking.

Similarly to the goals, the users' problems are expected to develop along the spectrum of their experience with survey data use. The outsiders' problems are very general. And while novice users' goals are associated with problems of low complexity, expert users are presented with very specific problems.

2.6.2 Measurement

Problems were measured with the question: "What are the main problems that you have encountered when finding or accessing survey data?" (Q16 Problems). The possible answers included different problems that had been mentioned repeatedly in the qualitative study by several interviewees. The listed answers range on a spectrum of very general problems ("I didn't know where to find data") to very specific problems ("Description and information on the data was incorrect"). Table 6 shows the answer options that were presented to the respondents, ordered from very general to very specific problems.

Table 6 Surveyed problems in ascending order of specificity

I didn't know where to find data.	Very general
I didn't know how to open or read the dataset.	
I didn't have the knowledge to understand the content of the dataset.	
I couldn't find data on my topic of interest.	
I couldn't find data on my population of interest.	
The data I found were too old.	
The data I found were of poor quality.	
Description or information on the data was insufficient.	
Description or information on the data was incorrect.	
I was denied access to data for legal or other reasons.	Very specific

2.7 Problem Solving

2.7.1 Operational Definition

According to the definition given by Wilson, information seeking behaviour basically is goal-oriented problem solving. For the present analysis, problem solving is defined as finding and applying strategies or measures to overcome *problematic situations* (Wersig and Windel 1985; Wilson 1999) when looking for data. Different problems require different strategies or measures of problem solving. Not every problem solving strategy can help with every problem. It is assumed here that having access to problem solving strategies that involve personal interaction is beneficiary. With regard to very specific problems (see Table 6) personal interaction may even be necessary for problem solving. On these grounds it is expected here that community involvement facilitates problem solving when looking for data. For example, problem-solving in this regard can be finding a previously unknown dataset or gaining access to restricted data through another community member. If information seeking behaviour is problem solving in its core, community involvement is key to successful data seeking.

2.7.2 Measurement

Problem solving was measured by asking the following question: "How do you deal with problems of finding and accessing survey data? Please indicate how important the following strategies of problem solving are for you." (Q17/18 Problem solving/closed) Respondents were asked to indicate the importance of each of these strategies on a scale from 1 to 5. The possible answers included strategies that require personal contacts (asking professors, colleagues, data specialists; participating in training or visiting a conference; finding help on

social media) as well as strategies that do not require these contacts (consulting documentation; conduct own survey; adjust research question).

2.8 Background

In addition to the questions on the key concepts, a range of questions was asked to produce general background variables, including age (Q25 Age), gender (Q26 Gender), country of residence (Q27 Country of residence), education (Q28 Degree), job status (Q29 Job status), stage of studies (Q30 Stage of studies), and sector or branch that the respondents work in or have last worked in (Q31/33 Sector/branch). Apart from being used as categorical variables, some of these questions also served as filters for the following questions. Further background variables were enquired with questions on the current or last job position (Q32/34 Current/last position); the professional experience, measured in years (Q35 Professional experience); and the research or study discipline (Q36 Discipline).

3. Data Collection

3.1 Questionnaire Preparation

The questionnaire was drafted in an offline document. After the first questionnaire evaluation steps (expert review), an online version of the questionnaire was created.

The online survey was created with the open source software 1KA OneKlick Survey[18] that is developed and provided by the University of Ljubljana, Slovenia (University of Ljubljana 2018). This software provides all functionalities that were needed for the present survey, such as comprehensive filter functionality (nesting and conditions), support of different languages, support of different devices, support of respondent debriefing and pre-testing, field monitoring and basic online analyses.

The survey was created in English. A first German translation was created after the first draft of the English pretesting version was ready.

3.2 Questionnaire Evaluation

In order to reduce measurement error, the questionnaire was thoroughly scrutinized and evaluated before entering the field (Groves et al. 2009). As current literature on

[18] https://www.1ka.si/d/en, accessed October 5, 2020.

questionnaire development suggests, a traditional field pretest (Groves et al. 2009) (or "conventional pretest" (Krosnick and Presser 2010, 296) or "pilot study" (Dillman, Smyth, and Christian 2009, 228)) may not be sufficient for questionnaire testing (Jacob, Heinz, and Décieux 2011). It has its strength in providing information such as average survey duration or proper functioning of filter questions. By provision of paradata (such as answering time per question) a field pretest can also provide an idea of comprehensibility of specific questions. However, these indications often give only vague evidence of the actual problem. Other methods of questionnaire evaluation such as expert reviews (Krosnick and Presser 2010) and respondent debriefings (Jacob, Heinz, and Décieux 2011; Krosnick and Presser 2010) are suited to provide more specific information on a questionnaire's weaknesses. This is why it is commonly recommended (Jacob, Heinz, and Décieux 2011; Krosnick and Presser 2010) to combine various evaluation methods to test the questionnaire before fielding. For the present study it was decided to evaluate and test the questionnaire draft through expert reviews, respondent debriefing, and field pretesting.

3.2.1 Expert Reviews of the Questionnaire

Expert reviews were obtained as a last step before preparing the online questionnaire. The primary goal of this step was to evaluate the questionnaire regarding its overall structure, question wordings, response alternatives and possible conceptual issues (cf. Groves et al. 2009). As it is usually recommended (Callegaro, Lozar Manfreda, and Vehovar 2015; Jacob, Heinz, and Décieux 2011; Krosnick and Presser 2010), the reviews were performed by experts in survey methodology or with extensive practical expertise and from the field of study (in this case, information behaviour). First drafts of the English questionnaire were presented to an online survey expert from a commercial market research institution (expert 1); a researcher with expertise in user studies (expert 2); a researcher with practical expertise in data collection and analysis (expert 3); and a researcher with expertise in survey methodology (expert 4). After each consultation, the questionnaire was revised before it was given to the next expert. Experts 1 to 3 gave their feedback on a paper version of the questionnaire, while expert 4 was presented with a first online version of the survey.

Expert 1 gave valuable feedback with regard to the order of response alternatives (randomized vs. non-randomized) and the treatment of no-options (e.g., "No answer."). Coming from outside information science or the social sciences, this expert could also give a

valuable assessment of question phrasing with regard to general comprehensibility, which helped to make the questionnaire more understandable and in parts less ambiguous. In line with their expertise in user studies, expert 2 in particular provided further valuable feedback on the question wording with regard to user behaviours such as finding or accessing data.

On grounds of practical expertise as a quantitative researcher, expert 3 helped to rephrase several items with direct reference to the social scientist's research process. For instance, this expert suggested grouping methodological skills (basic, advanced, expert) instead of surveying a long list of specific methods.

Expert 4 was presented with the first online version of the survey. Feedback from the first three experts had already been considered in this version. The feedback given by expert 4 in particular led to improvements with regard to the labelling of the rating scale points (values from "not important at all" to "very important" on a 5 point scale) as well as with regard to the visual layout in the software (for instance, the visual display of logos instead of survey names).

After the last expert review, a consolidated version of the online questionnaire was prepared for respondent debriefing. Additionally, a first German translation of the online questionnaire was prepared.

3.2.2 Respondent Debriefing

In general, respondent debriefings can be done in one of two alternative variants: the respondents are either asked to provide comments after they have completed the whole questionnaire; or they are asked to provide comments after each question (Jacob, Heinz, and Décieux 2011; Krosnick and Presser 2010). The most obvious advantage of the first alternative is that it allows for a realistic estimate of survey duration. On the downside, the respondents' comments on the questions will be less direct, especially if the questionnaire is long. Since it was planned to have a field pretest for estimation of the survey duration and since the questionnaire was quite long, it was decided to use the second alternative. The software used allowed for convenient online commenting with a text field for comments beneath each question.

By means of purposeful sampling, four participants were invited to complete the survey and provide comments on the questionnaire. Two of the participants are female, two are male.

Three participants were senior researchers (with doctoral degree) and one was a PhD researcher at the time. One of the participants is an English native speaker, three are German native speakers. Only the English native speaker was presented with the English version of the online questionnaire, the others were presented with the German version.

The respondent debriefing led to various small changes such as question wordings to clarify or disambiguate. Some larger changes were implemented as well. For example, in the question that enquires about knowledge of selected survey programmes (Q07/08 Known data/closed) some surveys were replaced according to the respondents' comments. All respondents have expert knowledge concerning the landscape of survey programmes. However, three specific surveys were unknown to 3 from 4 respondents (BIBB Erwerbstätigenbefragung and World Economic and Social Survey). Both were replaced by two popular Eurofound studies, as suggested by one respondent. Two respondents suggested adding studies from the Pew Research Center. When the following pretest showed that several respondents entered Pew studies in the free text box, the Pew American Trends Panel (Pew ATP) was also added to the selection.

The question on requirements when searching for data (Q12/13 Requirements/closed) was supplemented with an item on "availability of data free of charge". According to one respondent's suggestion, the question on problems that respondents had experienced in the past when looking for data (Q16 Problems) was specified to survey only the main problems, and reducing the answer options to 5 mentions ("Please give a maximum of 5 answers"). This change was expected to generate more significant results. Additionally, the "I never had problems" option was added and the originally preceding question on whether they had experienced problems was deleted.

Another item that was added is the "I conduct my own survey" option for the question on problem solving (Q17/18 Problem solving/closed). There was another problem revealed with regard to this question. One respondent expressed irritation about the labelling of the Likert scale. They would have expected a scale of frequency instead of importance. The original question "How do you usually deal with such problems? Please indicate how important the following strategies of problem solving are for you on a scale from 1 (not important at all) to 5 (very important)" would then have been changed to something like "How do you usually

deal with such problems? Please indicate how often you have used the following strategies of problem solving on a scale from 1 (never) to 5 (very often)". This seemed to be a good idea at first. However, the change would have been problematic with regard to the comparability of the question's items. For example, the items training/workshop participation or conference/event visit refer to measures that may not be used frequently but may have proved to be very important. For this reason, it was decided against changing of labels.

3.2.3 Field Pretest

Traditionally, field pretests are used to gain quantitative information about respondent behaviour and the process of questionnaire completion. The goal is to estimate, whether the survey will work well under realistic conditions (Groves et al. 2009). For the present study, the field pretest was designed according to standard conventions. With regard to the sampling this means that the survey was administered to respondents of the relevant population (Krosnick and Presser 2010), that is to say actual or potential survey data users. The sample included 16 respondents that had been purposefully sampled according to their professional background (Table 7).

Table 7 Respondents in pretest

Respondent no.	Gender	Language	Professional background
1	Male	German	Senior quantitative social scientist
2	Male	German	Senior quantitative social scientist
3	Female	German	Senior quantitative social scientist
4	Female	German	Senior quantitative economist
5	Female	German	Senior quantitative economist
6	Female	English	Senior quantitative social scientist
7	Male	German	Senior qualitative social scientist
8	Male	German	Senior historian
9	Male	German	Junior quantitative social scientist
10	Female	German	Junior quantitative social scientist
11	Female	German	Student of quantitative social sciences
12	Male	German	Student of quantitative social sciences
13	Male	German	Project manager (Market research)
14	Male	German	Freelancer (Market research)
15	Male	English	Market research professional
16	Female	German	Manager (PR and digital communications)

In the sample, there were 10 scientists, including 7 social scientists, 2 economists and 1 historian. Of these scientists, 8 were senior and 2 were junior. Eight researchers in the sample have quantitative orientation. Additionally, the sample included 1 female and 1 male student from the social sciences. The sample also included 4 respondents from outside academia who have all been working with survey data or survey results to differing degrees. In total, 9 male and 7 female respondents participated in the pretest.

One of the main reasons to conduct this pretest was to estimate the survey duration. On average, the respondents needed 12 minutes and 21 seconds to complete the survey. The calculation made by the software before the test had been 16 minutes and 28 seconds, which would have been too long. After looking at the results from the pretest, it was decided to give an estimate of 10 to 15 minutes in the invitation and consent form. Another takeaway from the survey duration evaluation was that the most time consuming questions (with durations > 1 minute) were the two Likert scale questions (the "requirements" and "problem solving" questions Q12 and Q17) and the yes/no table question (the "own contribution" question Q23). Questions of these formats are naturally more time consuming than less complex questions, which is why the comparably long duration of completing these questions was acceptable. All other questions turned out to have durations of less than one minute (ranging from 2.2 to 46.8 seconds), which indicates that there were no major problems to be expected in this area.

From looking at the pretest data, no technical problems became apparent. There were no unexpected missing values or inconsistencies and the conditions (filter questions) turned out to work as planned. However, looking more closely at the data, another possible issue stood out, relating to the question on knowledge of selected survey programmes (Q07/Q08 Known data/closed). As indicated above (0), pretest participants' comments in the free text field had also suggested that data from the Pew Research Center should be added. When investigating which other survey programme should be deleted instead, the decision was made to replace the GESIS panel. This decision was deduced from the pretest data, where the GESIS panel had as many mentions as the German General Social Survey ALLBUS, which seemed highly implausible given the download statistics of these two surveys. The datasets of both studies are available through the GESIS data catalogue DBK. Until 2018, all available GESIS Panel datasets were downloaded 455 times altogether, whereas the newest ALLBUS

dataset alone has been downloaded 5272 times (GESIS - Data Archive for the Social Sciences 2019). Hence, it seems more likely that respondents just read the familiar term GESIS and decided to click. So this measurement was saying more about awareness of GESIS than of this particular survey programme. On these grounds, it was decided to replace the GESIS panel item with an item for the Pew American Trends Panel (Pew ATP).

Even though the respondents had not been asked to give feedback on the survey questions, some of them chose to do so anyway and wrote back by e-mail. Some of the comments seemed quite significant and led to further amendments of the online questionnaire. For example, one of the mobile users indicated that the display of the survey programmes' logos in a grid seemed very crowded and should be optimized for mobile users (Q07). This suggestion was implemented by filtering mobile users to an alternative question that displayed the logos in on long row instead of a grid (Q08). Furthermore, two items were added according to the respondents' feedback: the question on where to find data (Q15 Seeking/sources) was supplemented by an item for "data archives"; and the question on problem solving (Q17/18 Problem solving/closed) was supplemented by the option "I conduct my own survey".

One problem that had come up in respondent debriefing was again mentioned by one of the pretest respondents: the labelling of the Likert scale for question Q17/18 Problem solving/closed: "How do you usually deal with problems of finding and accessing survey data? Please indicate how important the following strategies of problem solving are for you on a scale from 1 (not important at all) to 5 (very important)". The participant suggested replacing the importance labels with frequency labels (e.g., "not at all" to "frequently"). Because of the reasons given before (biased measurement of items), it was again decided not to make this change. Instead, the term "usually" was deleted from the question to reduce the association with frequency for the respondents. However, during the field phase, a participant wrote back that they were irritated by the scaling as well. In hindsight, this issue might have needed even more consideration and possibly not only a change of the labels, but also of the items.

Several minor changes, mainly in question or item phrasing, were made as a result to the feedback given by the pretest participants.

3.3 Field Phase

The survey was conducted between November 28, 2018 and January 7, 2019 (41 days). The invitations were sent between November 28, 2018 and December 10, 2019. A reminder was sent between December 12, 2018 and December 20, 2018. Before the reminder was sent, 709 invitees had completed the survey. After the reminder was sent, another 679 respondents completed the survey.

It was expected from the beginning, that the sample taken from the data catalogue users would probably contain a bias towards relatively highly educated and experienced researchers. In hindsight, this turned out to be the case (see below "D.4.2 Background: Education and Survey Data Literacy"). For this reason, the sample was supplemented by self-selected respondents that were recruited by intercept-sampling (Toepoel 2016) on the data catalogue's website. The intercept survey lasted from November 28, 2018 until December 10, 2018 (13 days). During this time, 70 respondents completed the survey via the pop up in the data catalogue.

Altogether, 1,458 surveys had been completed by the end of the field phase (Table 8).

Table 8 Completed surveys by method of recruitment

	Completed surveys from invited registered users	Completed surveys from intercept sampling	Total
Phase I (28 Nov – 11 Dec)	709	70	779
Phase II (12 Dec – 7 Jan)	679	--	679
Total	**1388**	**70**	**1458**

The registered users of the data catalogue were sent an e-mail that invited them to take part in a 10 to 15 minute survey that was designed to find out more about their experiences with data searching and data reuse in order to make data services more user friendly (see Annex 21). In the invitation, they were also informed about the underlying PhD project. Included were a link to the online survey and a link to an informed consent form (see Annex 20) that was also linked on the start page of the online survey (see Annex 24). The start page again provided short information on the goal of the survey and the underlying PhD project as well as links to the institutions involved. The consent form gave more information on the project

and on the principal investigator. Furthermore, it explained the planned data handling and processing. It informed the users that by clicking on "start survey" at the end of the start page they gave their consent for their data to be included in this study. The additionally sampled participants that had clicked on the pop up invitation in the data catalogue were directed to the same start page and were thus provided with the same information and consent form.

3.3.1 Survey Population and Response Rate

The data catalogue that was used to recruit the users is the DBK[19] from the GESIS Leibniz Institute for the Social Sciences in Germany. The catalogue provides access to datasets from approximately 6,000 national and international studies from the social sciences. Not all of the registered users of the DBK have agreed to being contacted by GESIS to receive further information upon registration. Deducting those user accounts without a respective agreement, the list comprised 19,006 names and e-mail addresses.

Not all 19,006 e-mail addresses could be used, because 181 of them contained invalid syntax (e.g. missing domain code). This means that 18,825 e-mail invitations could be sent out. Of these invitations, 1,987 could not be delivered ("bounces"). Furthermore, 111 invitations received automatic replies pointing to alternative e-mail addresses. The invitations could not be forwarded to these alternative e-mail addresses, because the agreement to be contacted was only valid for the addresses that were provided upon registration. In sum, 16,727 eligible registered DBK users actually received the invitation. Of these, 1,388 have completed the survey, 242 have partially completed the survey, and 15,097 refused to participate directly or shortly after having entered the survey.

The response rate was calculated based on the AAPOR standards (American Association of Public Opinion Research 2016) as depicted in Table 9. The response rate is 7.4 percent or 8.7 percent if the partially completed surveys are counted. These rates may seem quite low, but are not uncommon for a web survey (Callegaro, Lozar Manfreda, and Vehovar 2015).

[19] https://dbk.gesis.org/dbksearch/, accessed April 15, 2019. Meanwhile, the data catalogue has been included in the more general service https://search.gesis.org/, accessed October 5, 2020.

Table 9 Sample size and response rate

Total sample used:	18,825
Non contact:	1,987
Other (auto reply):	111
Total eligible users:	16,727
Complete interviews:	1,388
Partial interviews:	242
Refusal and break off:	15,097
Response rate (RR1)[20]:	7.4%
Response rate (RR2)[21]:	8.7%

In general, response rates in web surveys are lower than response rates from other survey modes. A meta study by Katja Lozar Manfreda et al. demonstrated that, on average, for web surveys, response rates are 11 percentage points lower than for other survey modes (Lozar Manfreda et al. 2008). Overall, response rates in web surveys vary broadly and seem to depend on many different factors. How these factors were addressed in this study is explained in the following paragraphs on recruiting.

3.3.2 Recruiting

When planning this study, several measures were taken to increase response. First of all, it was a list-based survey, which usually leads to higher response rates (Callegaro, Lozar Manfreda, and Vehovar 2015). Second, the survey was designed and conducted according to general recommendations on nonresponse. These recommendations refer to the following factors: the survey sponsor, topic salience, incentives, the contacting process, and the invitation format (cf. Callegaro, Lozar Manfreda, and Vehovar 2015).

Concerning the *survey sponsor*, literature recommends that trust and willingness to participate improve if a sponsor of legitimate authority is presented in the invitation or in the introductory page of the survey. In the present study, this sponsor was GESIS Leibniz Institute for the Social Sciences, which is the institution that hosts the data catalogue where the respondents had registered to download data. Furthermore, the Humboldt-Universität was mentioned, and weblinks to both institutions were provided. The supervisor of the

[20] RR1 = Response Rate 1 = "[…] the number of complete interviews divided by the number of interviews (complete plus partial) plus the number of non-interviews (refusal and break-off plus non-contacts plus others) plus all cases of unknown eligibility" (American Association of Public Opinion Research 2016, 61).
[21] RR2 = Response Rate 2; "counts partial interviews as respondents" (American Association of Public Opinion Research 2016, 61).

study at Humboldt was named in the consent form (Annex 19) as well as the data protection officer of GESIS (cf. Callegaro, Lozar Manfreda, and Vehovar 2015).

The factor of *topic salience* refers to the question whether the topic of the survey is important or relevant for the respondents (Callegaro, Lozar Manfreda, and Vehovar 2015). In the invitation (Annex 21), the survey was presented to the respondents as a chance to influence improvement of data services, which should be relevant for them, since they are all registered users of a specific data service.

Incentives have been shown to increase response rates in web surveys, but they are not completely unproblematic. For the present study it was decided not to use incentives. The main reason was that for populations with high education and income, incentives usually have a lower impact (Callegaro, Lozar Manfreda, and Vehovar 2015).

Regarding the *contacting process*, relevant factors that have been discussed in the literature are: day and time of sending invitations; number of contacts and their scheduling; length of the fieldwork period (Callegaro, Lozar Manfreda, and Vehovar 2015). If reminders are used, as it was done in this study, the day and time of sending are less relevant. Regarding the number of contacts, Callegaro et al. conclude from the literature that "the two-email setting (i.e. an invitation plus thank you note/reminder to all respondents) is enough. This setting was adopted in the present survey. All respondents were sent one reminder in the form of a thank you note that also asked them to complete the survey if they had not already done so (Annex 22). Concerning the length of the fieldwork period, it was decided not to make it too short. Since there was no immediate reason to finish as soon as possible, the survey was kept open for six weeks (41 days). Not surprisingly, the vast majority of surveys were completed in the first four weeks of this period and participation decreased abruptly afterwards (Figure 15).

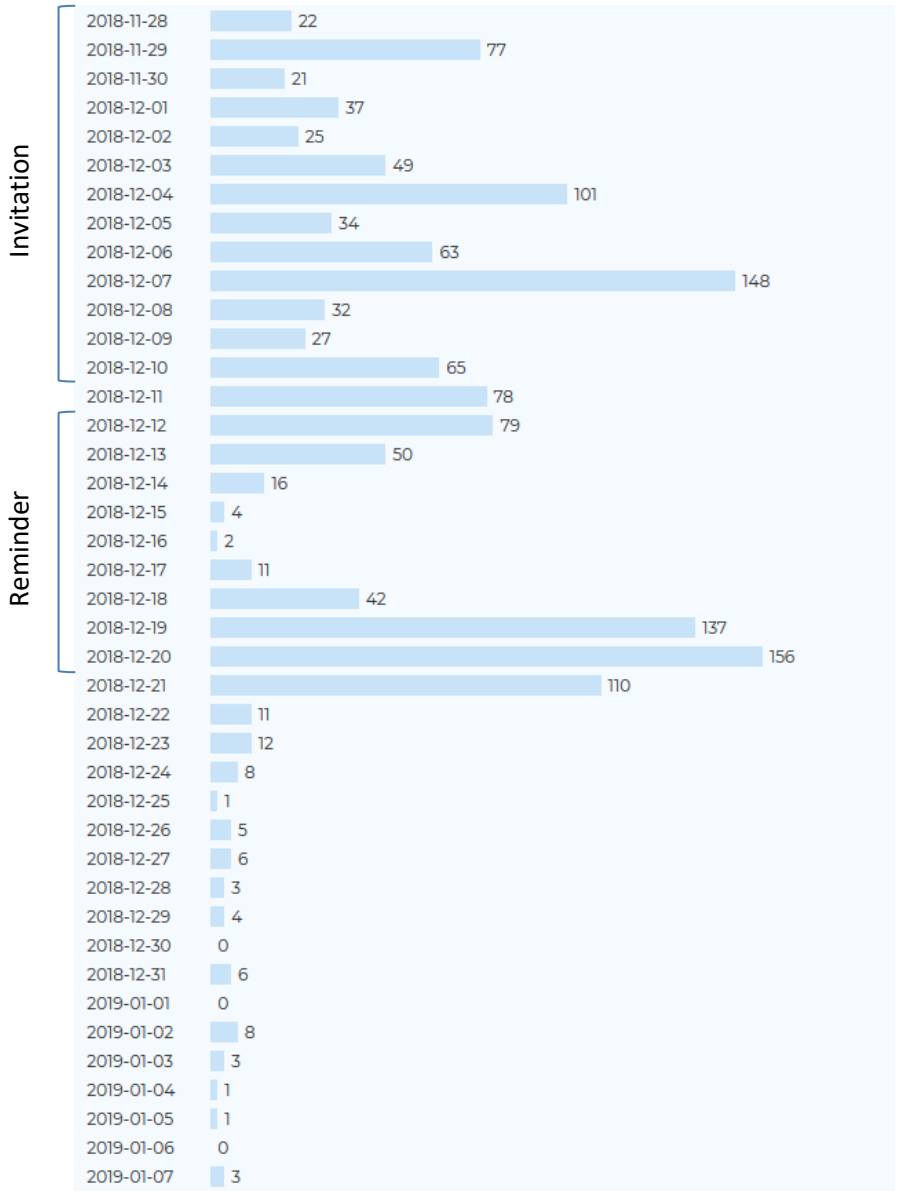


Figure 15 Completed surveys per day during the field phase (image produced by 1KA OneKlick Survey)

Concerning the *invitation format* certain specifics should be considered when conducting a web survey (Callegaro, Lozar Manfreda, and Vehovar 2015). The e-mail invitation to the survey is supposed to "provide basic information, legitimacy, instructions and motivation" (Callegaro, Lozar Manfreda, and Vehovar 2015, 154). All these aspects were addressed when designing the invitation (Annex 21). The description of the survey project was kept concise but informative. The background motivation of the study (improvement of services as well as PhD research) was clearly stated. The invitations were sent from a GESIS affiliated e-mail account (survey.datasearch@gesis.org) to establish trust. At the end of the invitation text,

recipients were informed about the source of their contact data and the legitimacy of being contacted. The associated institutions were named (GESIS and Humboldt Universität) and contact data (postal address and personal e-mail address) of the principal investigator were included in the text. A link with further information on data protection at GESIS was also included. Links to the survey (in English and German) and short instructions as well as information on the length of the survey were provided. The issue of motivation was addressed by a short header that included the two main topics that recipients were expected to relate to (data search and data use). Recipients were informed that their contribution was valued as support for infrastructure improvement and could have broad impact on infrastructure development. In order to keep the invitation as short as possible, detailed information on the research project, data handling, and data processing was not included in the text. This information was given in a separate consent form (Annex 19), and the link to this consent form was provided in the invitation and again on the start page of the survey.

Finally, a very important factor regarding readiness and motivation to participate in a survey is personalisation of the invitation. Therefore, all recipients were addressed with their full name in the salutation.

In sum, a wide range of measures were taken to ensure broad response.

3.3.3 Complementary Sample

A pop-up window on the website invited users to participate in a survey on data searching and data use (see Annex 23). Upon seeing the pop-up for the first time, people were informed to enable cookies, so they would not be presented the pop-up again if they had already declined participation. This pop-up was online for 13 days. The sample could then be enriched by some less experienced users (mainly students), but in total, only 70 valid cases resulted from this source during the time period of 13 days, as opposed to 1,388 valid cases that were sampled during the 41 days of inviting registered users.

3.3.4 Sample Size

Even though the response rate of the survey was rather low, the sample size was high enough to make the intended analyses with regard to the target population. The size of the population, defined here as the registered users of the GESIS data catalogue, was 16,727

(calculated from 18,825 listed e-mail addresses minus 1,987 bounces minus 111 re-directs, cf. Table 9). If the confidence level is set at 95 percent certainty and the confidence interval is at 5 percentage points, 376 respondents are needed. The standard error on these grounds would be at 5.11 percent of the estimate.[22] The analyses that were made in this study and are presented in the following paragraphs are all based on estimates with more than 376 respondents. The smallest number of individuals in an estimate that is made in the following is 478. For this number, the standard error is at 4.5 percent of the estimate, if the confidence level is set at 95 percent and the confidence interval is at 5 percentage points.

3.4 Data Processing

After the field phase, the survey dataset was downloaded using the export functionality provided by the 1KA OneKlick software (University of Ljubljana 2018). The download is only provided in .sav format and since the analyses were done with Stata/SE 15.1 (StataCorp 2017), a .dta file was created using the Stat/Transfer data conversion software (Circle Systems 2015).

Afterwards, data cleaning and recoding was done using Stata, which was also used for analysis.

3.4.1 Data Cleaning

To prepare the raw Stata dataset for analysis, several data cleaning steps were necessary. All these steps were performed by means of a Stata do-file created for this purpose. The do-file performs the following steps on the raw data file:

- It deletes invalid (empty) units from the dataset. These stem from participants who entered but did not complete the survey.
- It changes variable names (Q01, Q02, etc.) to more descriptive names (age, gender, etc.)
- It reassigns country codes. This step was necessary, because the list of countries presented to the respondents sorts differently in the two survey languages English and German. The resulting country codes in the variable "country" were running numbers that pointed to different countries depending on the chosen questionnaire language. The

[22] The sample sizes were calculated with the sample size calculator provided by the Australian Bureau of Statistics, URL: https://www.abs.gov.au/websitedbs/D3310114.nsf/home/Sample+Size+Calculator, accessed October 5, 2020.

do-file creates a new variable "residence" with unified country codes created from "country", depending on "language".

- It creates combined versions of variables that had to be collected separately depending on the type of device that was used to complete the survey (PC and mobile devices). The reason for the occurrence of these split variables is that questions with many items, with Likert scales, or with graphic enhancements had to be designed in mobile friendly versions. As a result, the program created different variables for the same question depending on the device that was used. These variables are combined in new unified variables by the do-file and afterwards all variables that contain values for either pc or mobile devices are dropped.
- It creates a new variable "methods" that summarizes the methodological skill indicated in the "meth*" variables (three variables that indicate basic, advanced, and expert methodological skills) in a scale from 1 to 3.
- It creates new variables "branch" and "position" that combine the values regardless of whether branches and positions are current or past (in case of current unemployment or retirement).
- It creates a new variable "discipline" that contains one discipline code per case.
- It replaces numeric codes (-1, -2, … -99) for missing values by corresponding Stata codes (.a, .b, … .i).
- It adds proper labels to all variables.
- It adds value labels to all variables using variable containers.
- It deletes irrelevant system variables ("invitation", "lurker", and "code").
- It sorts variables in their order of appearance in the questionnaire.
- It deletes remaining irrelevant cases (value of variable relevance is 0).

3.4.2 Data Recoding

The questionnaire contained several questions with an open answer option. For each of these questions, the open answers were exported into an Excel file for recoding. As it turned out, several respondents had given answers in the open text fields that had actually been possible answer options. In these cases, the answers were recoded using the existing codes. This work was again done with a Stata do-file that was created for this purpose.

Those contributions that indicated answers that had indeed not been at choice in the answer options were treated as new answer categories. They were assigned numeric codes and short labels in the Excel sheet. Even though some of these categories were found in multiple cases, they were not added to the dataset as new answer categories. The main reason is that even more respondents might have indicated these categories had they been answer options in the first place. The mentions of these categories are therefore incomplete and can only be regarded as additional relevant information that may be used for future research. The most frequently mentioned categories (more than 10 occurrences) are listed in Table 10 together with the respective question.

Table 10 Most mentioned categories in open answers

Question	New category from open answers	Occurrences
For what purposes did you use survey data in the past two years?	- **University assignment**	15
	- **Consulting**	12
Where do you know these survey programmes from?	- **Personal contribution to the survey programme** (other than being principal investigator)	13
	- **General research**	15
When searching for these data, how important were each of the following requirements? … I had other important requirements:	- **Accessibility** (e.g. technical barriers, costs)	15
What are the main problems that you have encountered when finding or accessing survey data?	- **Accessibility** (e.g. technical barriers, costs)	37
How have you shared your survey datasets? Please think of any survey data that you have shared in the past. I have … shared survey data in another way:	- Upon **personal request**	11
	- Shared with **students**	11
Some people who are working with survey data contribute to the creation, improvement, or dissemination of survey data for reuse in some way or another. Have you ever engaged in one or more of the following activities?	- Contributed to **data infrastructure**	11

The cleaned and recoded dataset can be downloaded for scientific reuse from the GESIS data archive (Friedrich 2020b).

4. Description of the Sample

4.1 Basic Demographics

From all participants in the study, 1.458 have completed the survey. 943 (64.68 %) participants chose to complete the survey in English and 515 (35.32 %) answered the German version of the survey.

The following paragraphs describe the sample by the indicated demographics, starting with the basic demographic variables age, gender, and residence.

All surveyed age groups were present in the sample (Figure 16).

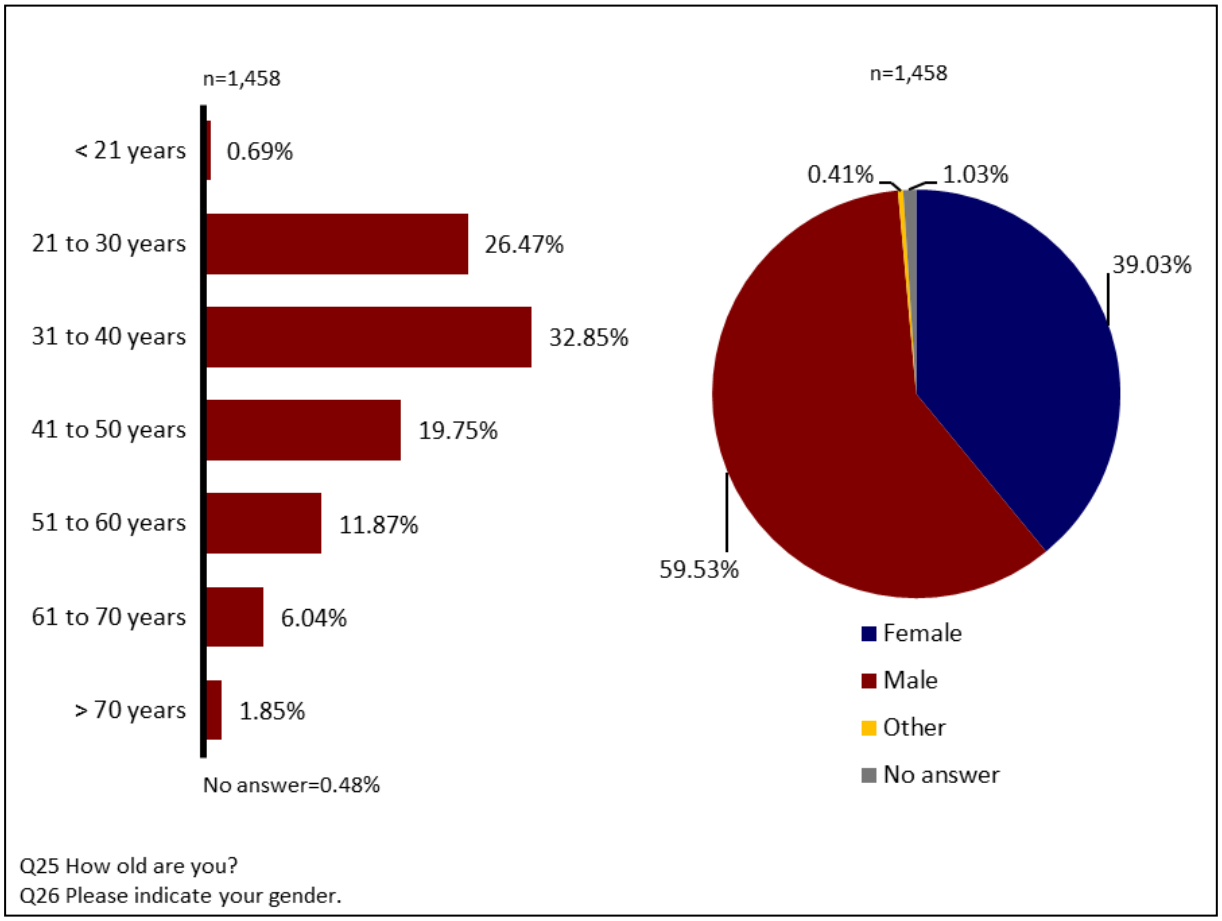


Figure 16 Age groups and gender distribution

The largest age group are the 31 to 40 year olds, who make up almost one third of the sample (32.85%). The second largest group are the 21 to 30 year olds (26.47%), followed by the 41 to 50 year olds (19.75%). Almost 80 percent of the sample are between 21 and 50 years old. The gender distribution seems uneven at first, with only 39.03 percent women in the sample as opposed to 59.53% men. However, given the underrepresentation of women in science, this distribution is less surprising. For instance, in Germany, only 29.3 percent of post-doctoral degrees (*Habilitation*) were obtained by women in 2017 (Statistisches Bundesamt 2018). In the field of law, economics and social sciences, the share of women who received a post-doctoral degree was at 35.6 percent.

Table 11 What is your current country of residence? (Countries with more than 10 respondents)

	Residence			Residence	
	(n)	(%)		(n)	(%)
Germany	454	31.14	Turkey	18	1.23
United States of America	101	6.93	Chile	17	1.17
Italy	81	5.56	Denmark	17	1.17
Spain	60	4.12	Poland	17	1.17
Austria	49	3.36	Russian Federation	16	1.10
United Kingdom	46	3.16	Croatia	15	1.03
Netherlands	40	2.74	Norway	15	1.03
Japan	37	2.54	Finland	14	0.96
Sweden	29	1.99	Ireland	13	0.89
China	29	1.99	Latvia	13	0.89
Portugal	26	1.78	Hungary	12	0.82
Greece	24	1.65	Slovakia	12	0.82
South Korea	24	1.65	Canada	11	0.75
Switzerland	23	1.58	Czechia	11	0.75
Belgium	22	1.51	Other countries*	150	10.29
France	22	1.51	No answer	19	1.30
Romania	21	1.44	Total	1,458	100

* All mentioned countries with n ≤ 10
Q27 What is your current country of residence?

A more interesting demographic characteristic of the sample turned out to be the distribution by country of residence. Even though all sampled participants are clients of a German research data provider, respondents who reside in Germany only make up about one third of the sample (31.14 %) (Table 11). Possibly because of the considerable number of international studies that are available through the data catalogue, two-thirds of the sample

are international. Notably, the vast majority of respondents (77.29%) reside in Europe (Figure 17). About 10 percent of the respondents reside in Asia (9.67%), and about 8 percent in North America (8.09%).

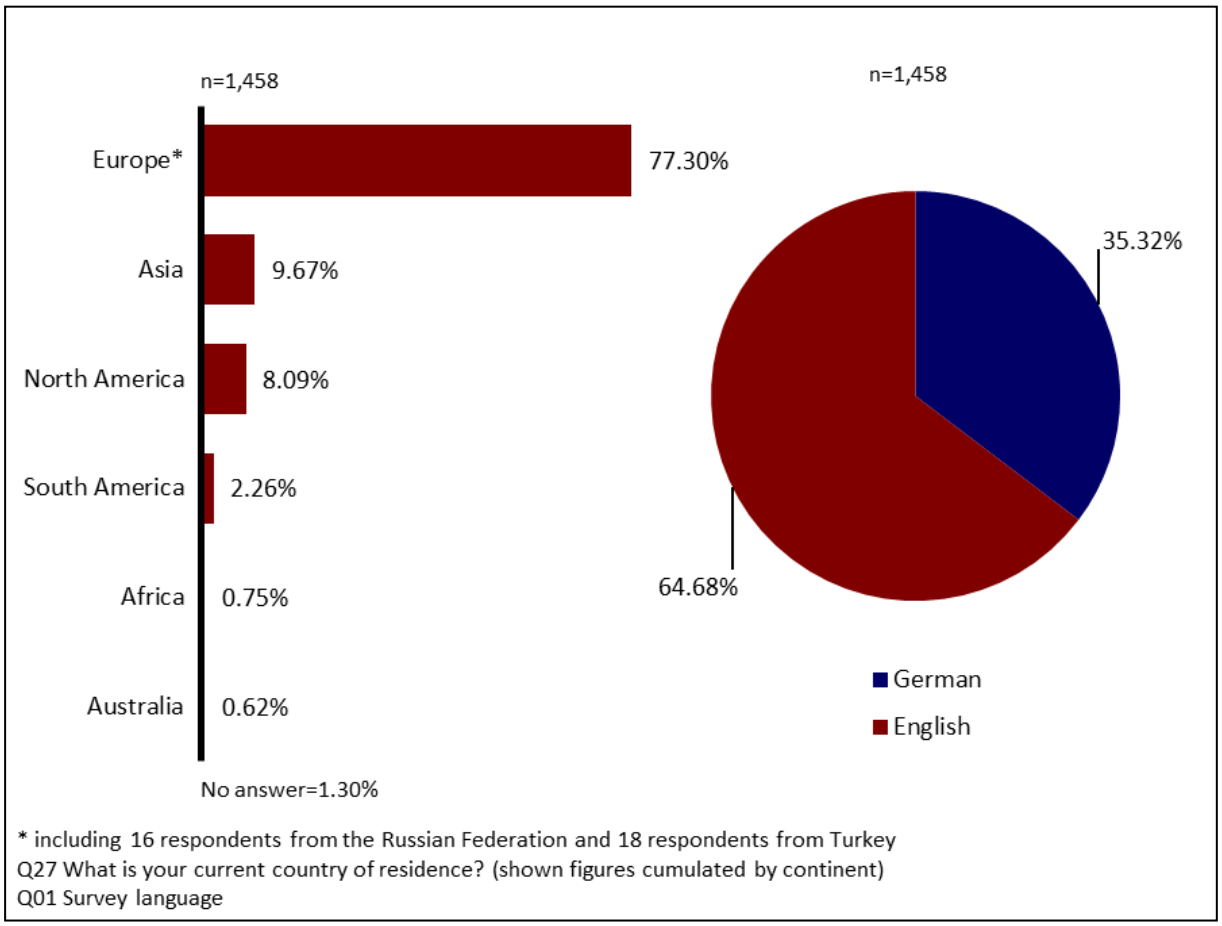


Figure 17 Residence by continent; chosen survey language

4.2 Background: Education and Survey Data Literacy

Beyond the basic demographics age, gender, and residence, the questions on education and employment allow for a first estimation of how experienced and qualified the surveyed data users are. Most notably, almost 40 percent of the respondents indicated that they had a doctoral degree (38.89%) (Figure 18). Together with those respondents who have a postdoctoral degree (11.59%), which is only available in some countries, the users with a doctoral or higher degree make up about half of the sample (50.48%).

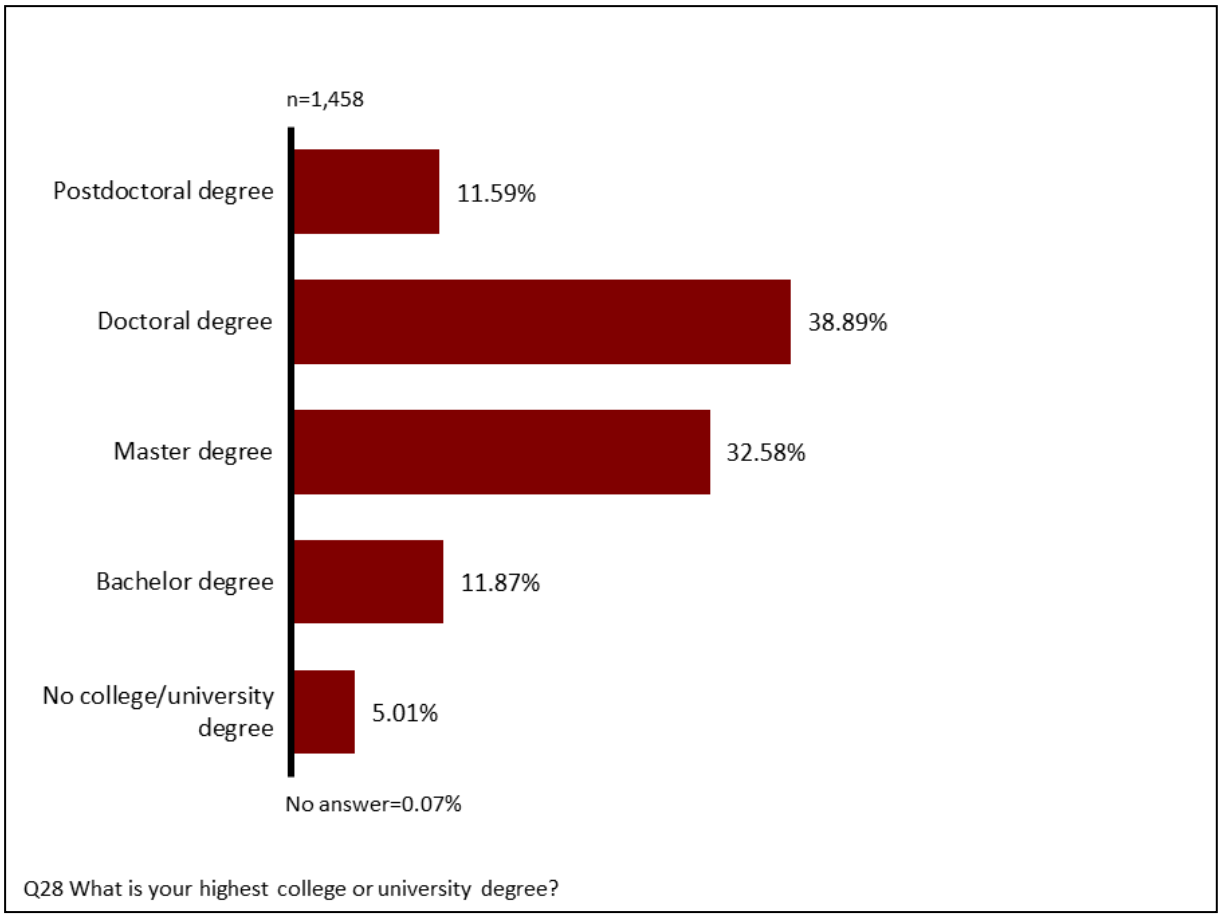


Figure 18 Highest college or university degree

In comparison, only about 20 percent of the respondents indicated being fulltime students (20.15%) and from these only 22.41 percent (67 individuals) replied they were bachelor students. More than 75 percent of the surveyed students indicated being master or PhD students (Figure 19). The vast majority of the non-students in the sample work in research, science or technology. Around 70 percent of the respondents have indicated working in these sectors.

From looking only at these distributions, it is to be expected that the general level of experience and research proficiency in the sample is very high. This impression is backed by the self-estimated methodological skills that the respondents have indicated (more than 80% replied that they had used expert methods to analyse survey data, see Figure 22).

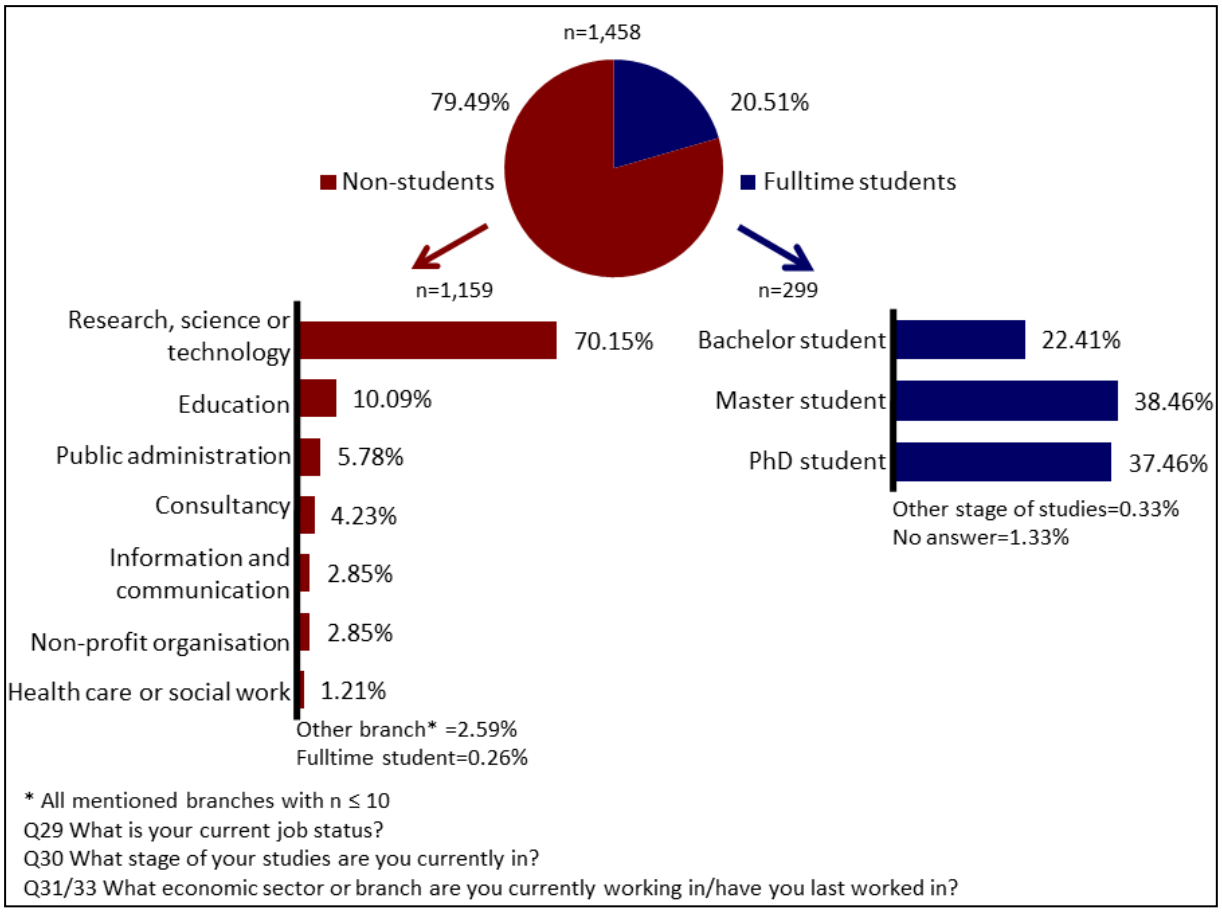


Figure 19 Economic branch and stage of studies

Not only do the respondents have very high levels of education, but those who are working in research and technology (813 individuals[23], 70.15% of the sample) tend to be in high positions: 41.02 percent (331 respondents) indicated being university or college professors (Table 12). Together with other lecturers (8.18%) and senior researchers (27.14%) they make up 76.34 percent of the respondents who work in research and technology.

[23] Originally, 807 respondents had indicated research, science, or technology as their branch of work; after recoding of the open answers, this number was corrected to 813. The 5 additional respondents in this category have not been asked Q32/34, so there is no data on their current position in research, science, or technology.

Table 12 Current or last position in research and technology

	Position	
	(n)	(%)
University/college professor	331	41.02
Lecturer at university/college	66	8.18
Senior researcher or postdoc	219	27.14
Junior researcher or PhD student	145	17.97
Research assistant (bachelor degree)	15	1.86
Librarian	4	0.50
Administrator	4	0.50
Other position	19	2.35
No answer	4	0.50
Total	807	100

Q32/34 What is/was your position with your current/last employer?

The results of self-reported experience in the present occupation or earlier similar occupations again back up the impression that the sample population is very experienced; on average, the respondents have been working in their current occupation for about thirteen years (the arithmetic mean is 12.9708 years with a standard deviation of 11.11916) (Table 13).

Table 13 How long have you been in your job or in similar jobs that you have had before?

	N	Mean	Std. deviation	Min	Max
Years in current or similar jobs	1096	12.9708	11.11916	0	57

Finally, the potentially high expertise in survey analysis is also reflected in the high number of respondents that study or work in academic fields that typically use survey data, such as

social sciences, economics, education, and psychology (979 respondents, 67.15% of the total sample) (Table 14).

Table 14 What is your field of research/ field of study in Humanities and Social Sciences?

	Field (n)	Field (%)
Social Sciences	735	69.47
Economics	168	15.88
Psychology	51	4.82
Humanities and Social Sciences (general)	42	3.97
Educational Research	25	2.43
History	9	0.85
Social and Cultural Anthropology ...	9	0.85
Jurisprudence	6	0.57
Theology	3	0.28
Linguistics	3	0.28
Fine Arts, Music, Theatre and Media Studies	2	0.19
Library and Information Science	2	0.19
Philosophy	1	0.10
Ancient Cultures	0	0.00
Other	2	0.19
Total	1,058	100

Q37 What is your field of research / field of study in Humanities and Social Sciences?

To assess the respondents' experience with survey data research (survey data literacy), specific questions regarding their practice with data use, data analysis and use of methodology were included in the survey. This part of the survey started with a general filter question: Have you ever used survey data for your work or for your studies? This question was meant to sort out respondents who had no experience with survey data analysis. Those respondents were asked only a subset of questions that were intended to find out to what degree they were aware of survey data or surveys at all. Only very few participants indicated that they had never used survey data or any other research data for their work or for their studies: 4.25% (62 respondents) (Figure 20). Almost two percent (28 respondents) reported that they had never used survey data, but other research data. Twenty-one respondents gave further details as to what other research data they had used. Statistics and qualitative

data were the most frequently mentioned research data types. The vast majority of participants (1368 individuals, 93.83% of the sample) confirmed that they had used survey data for their work or for their studies. Given that all participants were sampled from users of a survey data catalogue, these numbers are not surprising.

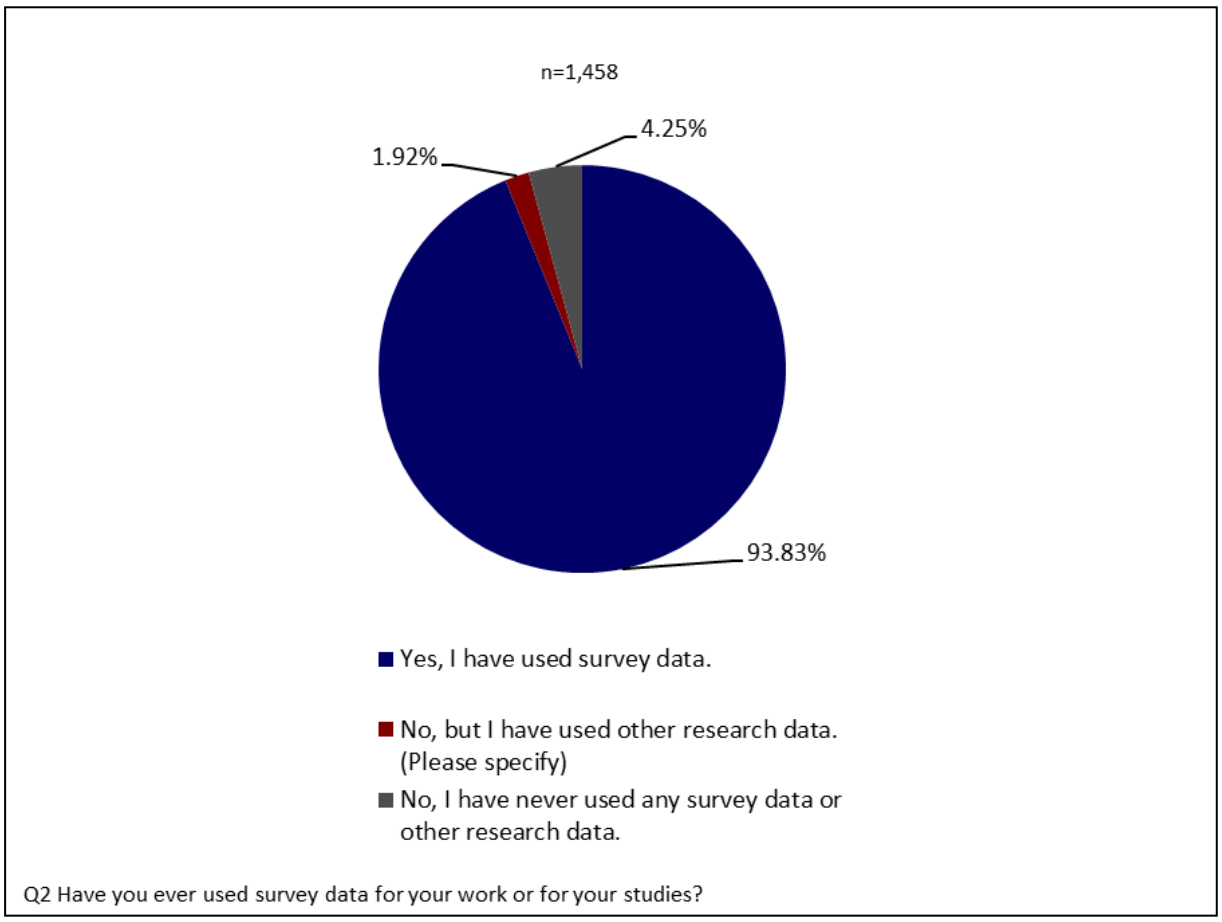


Figure 20 Use of survey data for work or for studies

Respondents who had already used survey data were further asked whether they had already performed statistical analyses with these data. Of the 1345 participants who answered this question, a total of 1303 respondents (96.88%) indicated that they had performed statistical analyses before, including 262 participants (19.48%) who stated that they had done this kind of analyses only once or twice and 1041 (77.40%) who had used statistical analysis more often (Figure 21). Only 42 respondents (3.12%) had never performed statistical analyses with survey data.

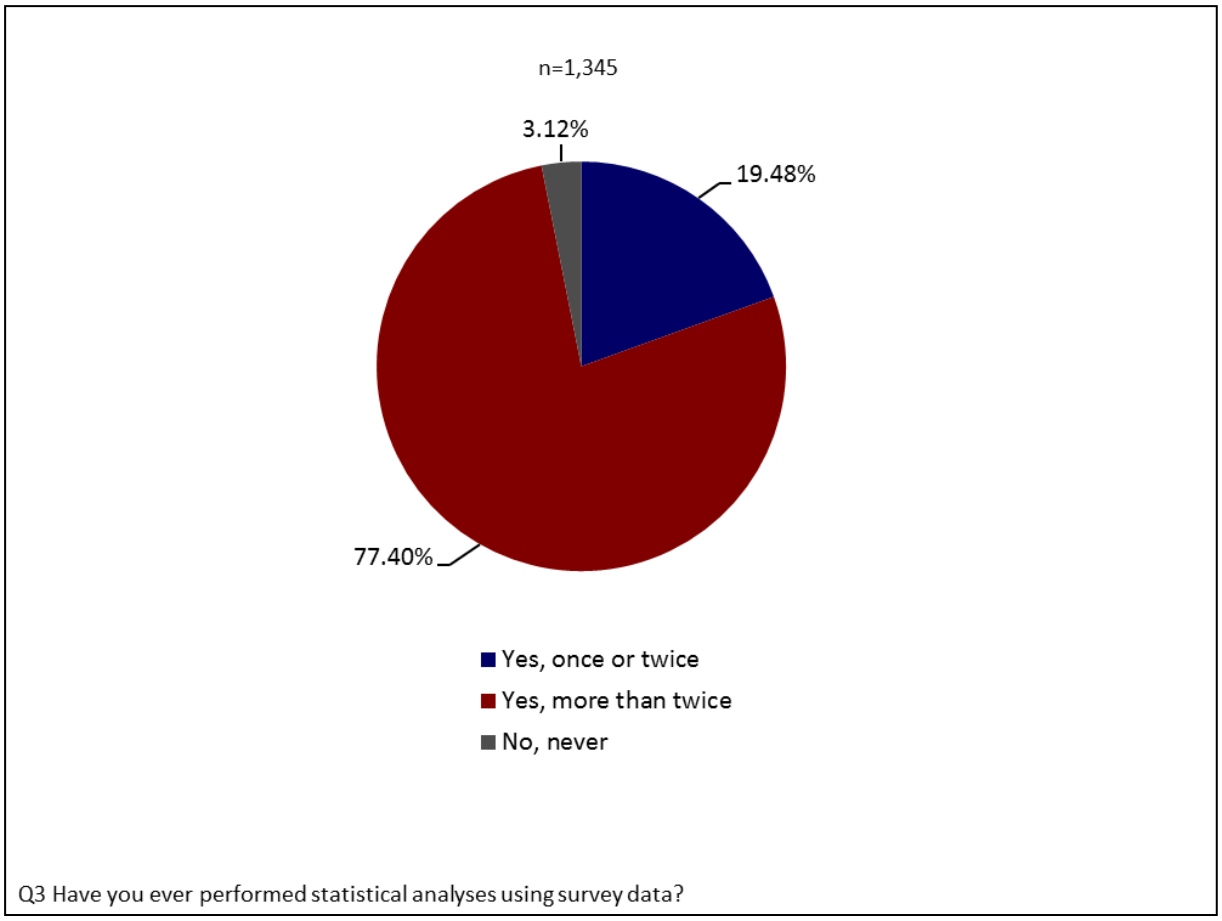


Figure 21 Statistical analyses with survey data

Out of those who have performed statistical analyses, a majority of 82.27 percent (1072 respondents) indicated to having used expert methods of analysis, such as multiple regression or other multivariate analyses (Figure 22). While 11.44 percent (149 respondents) have used advanced methods such as cross tabulation or other bivariate analyses, only 6.29 percent (82) have used basic methods such as counting, frequencies, distributions or other univariate analyses.

Furthermore, 1,345 respondents who had indicated that they had used survey data in the past were asked, whether they had ever conducted a survey and produced survey data on their own or together with other people. Nearly three-fourths, or 1,001 respondents (74.42%) confirmed that they had collected data in the past. All these numbers point to high levels of experience with data use and analysis in the sample.

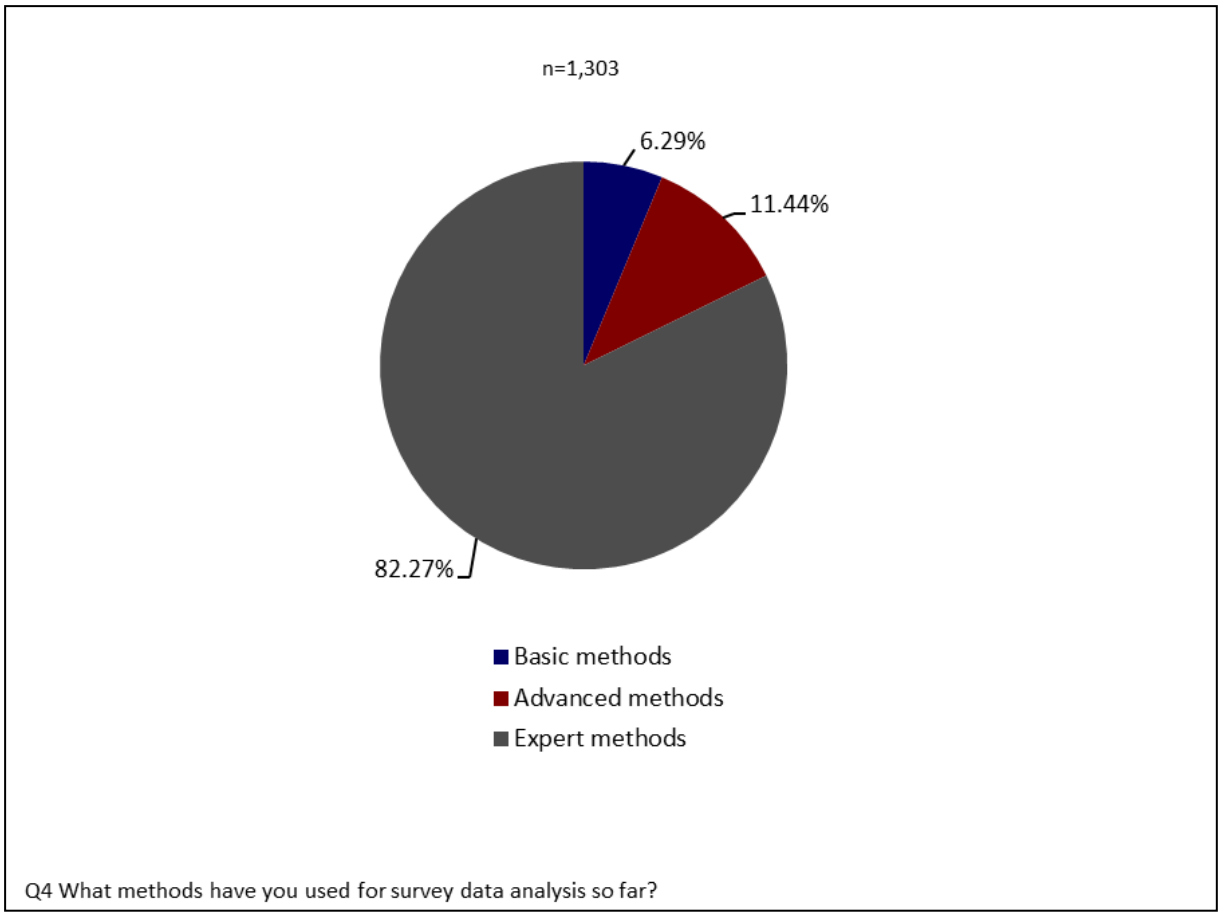


Figure 22 Methods of survey data analysis

5. Development of the Experience Index and the Community Involvement Scale

An experience index and a community involvement scale were created to measure two core concepts that were needed to analyse some of the hypotheses.

5.1 The Experience Index

This experience index was created as a formative index. For this kind of index, the values of multiple indicators are added up for each case (Schnell, Hill, and Esser 2013). The sum of the indicator values represents the index value for each case. In the case of the experience index, four indicators with values from 0 to 3 were combined in a formative index ranging from 0 to 12. For the creation of the experience index that measures experience in survey data analysis, four indicators were combined: work amount, work type, work type, and work knowledge. These indicators are based on measurements of work experience (Quińones, Ford, and Teachout 1995). The formation of the experience index is laid out in detail in the following paragraphs.

Experience, as it is understood here, refers to the accumulated knowledge and practice in survey data analysis. This knowledge and practice was measured in different *Measurement Modes*, as suggested for measurements of work experience by Quińones, Ford and Teachout (1995). Quinones et al. introduce the modes of *amount*, *time*, and *type of work* and suggest applying these modes to a *Level of Specificity* that is relevant for the research question (Quińones, Ford, and Teachout 1995). They present three levels of specificity: *organisation*, *job*, and *task* (Figure 23).

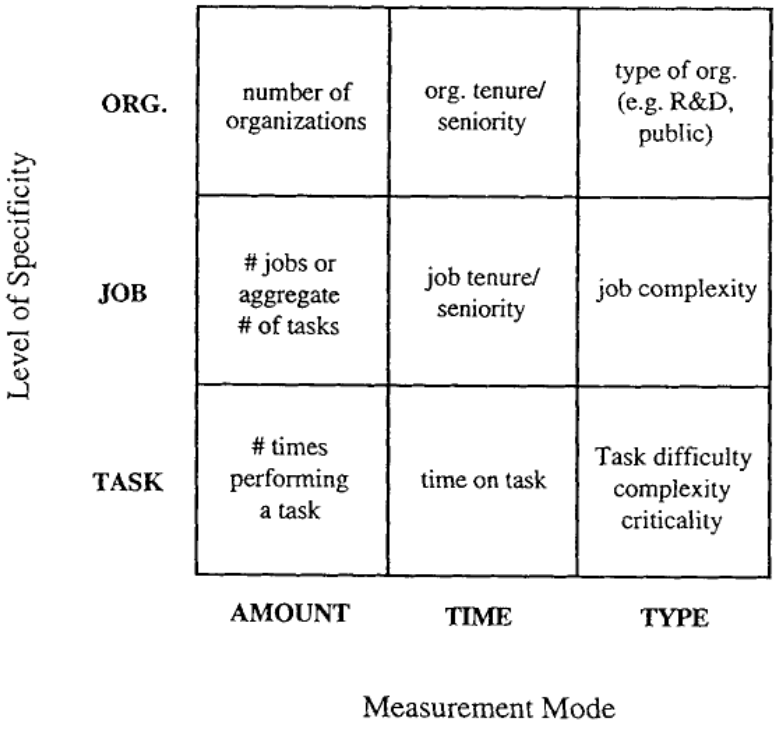


Figure 23 A Conceptual Framework of Work Experience Measures (Quinones et al. 1995, 892)

For the measurement of experience in survey data analysis, the specificity ranges between the task level and the job level, because survey data analysis is a more or less complex conglomeration of various tasks that occurs in a broad variety of job contexts. For this reason, it was decided to introduce the *work* level as an in-between *Level of Specificity*, ranging between the job level and the task level.

Furthermore, regarding the *Measurement Mode*, the conceptual framework by Quińones et al. does not include dimensions of work *knowledge*, which may be a less critical dimension in

many kinds of work, but with regard to work in science or research, it seems essential. Consequently, the new dimension of *work knowledge* was added to the experience index as described below. Along the now four measurement modes of *amount*, *time*, *type*, and *knowledge* the dimensions of work in the present context can be described and were measured as described below and as depicted in Figure 24.

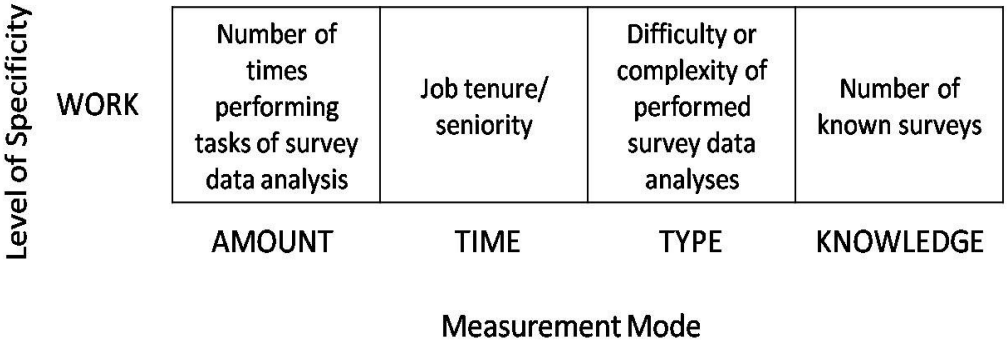


Figure 24 Conceptual framework of work experience in data analysis

Work amount*: Number of times performing tasks of survey data analysis.* The amount of work in survey data analysis was measured by combination of two questions: Q03 "Have you ever performed statistical analyses using survey data?" (answer options: never/once or twice/more than twice, distributions in Figure 21) and Q20 "Have you ever conducted a survey and produced survey data (either on your own or together with other people)?" (answer options: yes/no). The resulting variable provides values on a scale from 0 ("no experience") to 3 ("much experience"). Respondents who have indicated "never" and "no" in questions Q03 and Q20 have "no experience" (index value = 0). Only if they have indicated either "once or twice" in Q03 or "yes" in Q20, their value on the experience index is 1. Value 2 is reached by respondents who have answered "more than twice" in Q03 and "no" in Q20 as well as by those who have answered "once or twice" in Q03 and "yes" in Q20. Only those respondents who have answered "more than twice" in Q03 and "yes" in Q20 reach 3 points ("much experience") on the experience index.

Work time: *Job tenure/ seniority*, as proposed for the level "job" in the conceptual framework (Figure 23). The most useful background variable for this measurement turned out to be the question on the highest college/university degree (Q28). Other potential

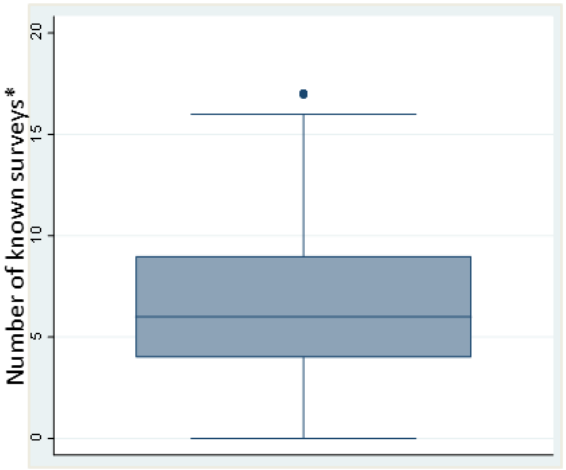
candidates were Q29 ("What stage of your studies are you currently in?"); Q32 ("What is your position with your current employer?"); Q35 ("How long have you been in your job?"). However, depending on their job status, these questions were not presented to all participants (students, employees, etc.). Mainly for this reason, it was decided to use the measurement of educational degree (distributions in Figure 18). The variable provides values on a scale from 0 ("No college or university degree") to 3 ("Doctoral or postdoctoral degree").

Work type: *Difficulty or complexity of performed survey data analyses*. This dimension was measured with question Q04 "What methods have you used for survey data analysis so far?" This variable provides the values 1 ("basic methods"), 2 ("advanced methods") and 3 ("expert methods"). The distribution of this question are depicted in Figure 22. Respondents with no experience with survey data analysis have a missing value here that was counted as 0 when forming the index.

Work knowledge: *Number of known surveys*. This dimension was measured by question Q07/08 "Have you ever heard of the following survey?" The respondents were presented with a list of 25 selected survey programmes, assuming that people with more experience in survey research should know more surveys than people with less experience. The preselection of the 25 surveys is described above in subchapter D.2.2.2. For assessment of knowledge about survey programmes, a new variable had to be created based on the surveys variable produced by Q07/08. The main reason for this is that the 25 selected surveys in this question are biased towards German surveys and would therefore lead to non-Germans scoring lower on survey programme knowledge on average. Hence, the adjusted variable surveys10 does not include surveys that are known by less than 10 percent of either English or German respondents. Excluded surveys are: British Social Attitudes; German Longitudinal Election Study; CILS4EU; German Internet Panel; National Educational Panel Study; Pairfam; Shell Youth Study; and GMF (Gruppenbezogene Menschenfeindlichkeit). The adjusted variable includes values for 17 of the initial 25 surveys and thus ranges from 0 (respondent never heard of any of these surveys) to 17 (respondents heard of all these surveys). In order to give this variable the same weight as the other three variables in the index, the values were aggregated to a range from 0 (heard of 0 surveys) over 1 (heard of 1 to 3 surveys) and 2 (heard of 4 to 10 surveys) to 3 (heard of 11 to 17

surveys). These boundaries were chosen by calculating the mean (6.607682) and subtracting (for the lower boundary of the middle category) or adding (for the upper boundary of the middle category) the standard deviation (3.751147). The table in Figure 25 shows the distributions of this variable. The borders are visualized in the box plot (Figure 25).

	Number of known surveys*	
	(n)	(%)
0 surveys	43	2.95
1 to 5 surveys	305	20.92
6 to 11 surveys	865	59.33
12 to 17 surveys	245	16.80
Total	1,458	100



* Includes only those surveys that were known by at least 10% of respondents
Q7 Have you ever heard of the following survey?

Figure 25 Number of known surveys (table and box plot)

When constructing a formative index, it is important to combine items that contribute equally and as independently as possible to the measured construct (in this case: experience) (Schnell, Hill, and Esser 2013). At the same time, the items are supposed to correlate positively (Latcheva and Davidov 2014). This is why the correlation of the items that were used to form the index was tested beforehand. The correlation analysis revealed that the items are all moderately positively correlated, which is acceptable when creating a formative index (Latcheva and Davidov 2014).

The resulting experience index ranges from 0 (no experience) to 12 (high experience). The distribution of the experience index is depicted in Figure 26.

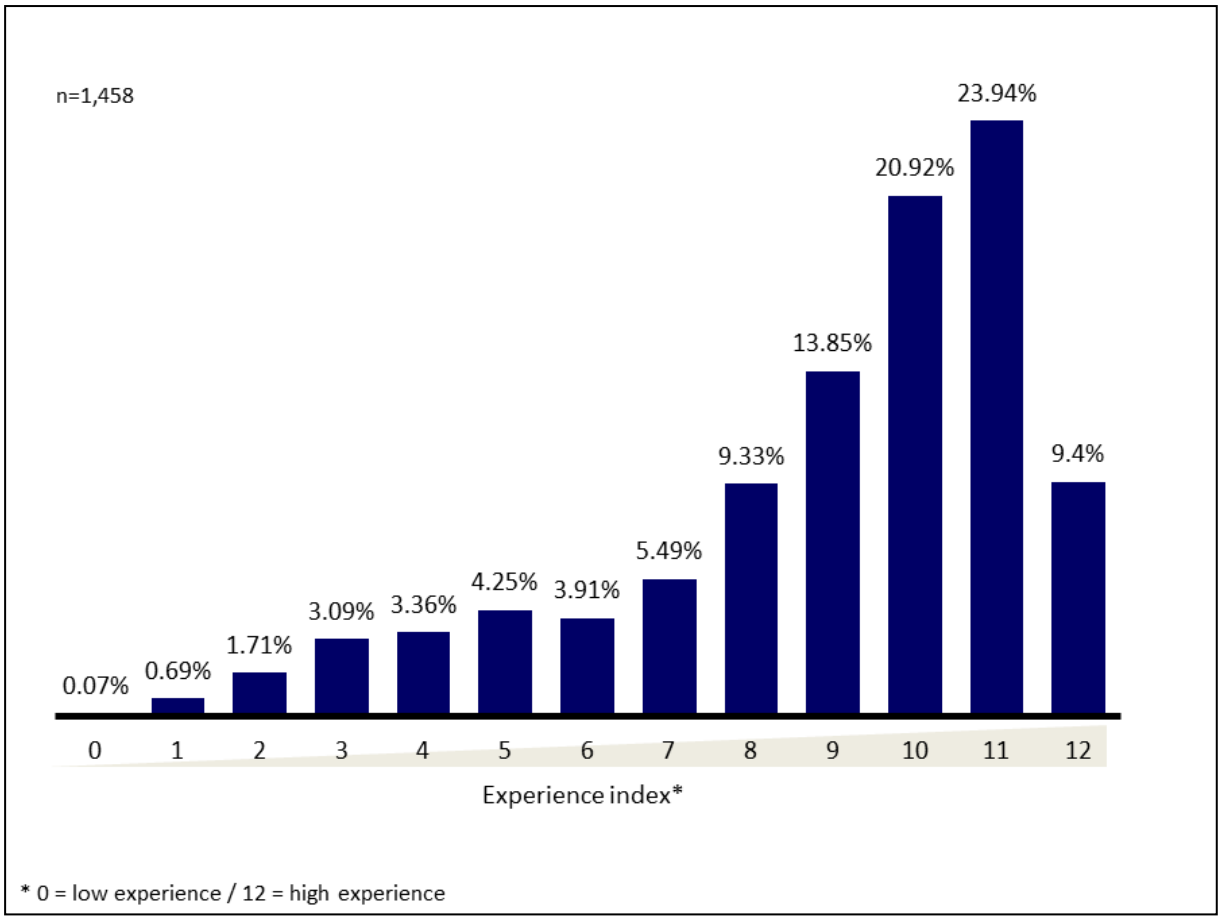


Figure 26 Distribution of experience index

5.2 The Community Involvement Scale

In order to assess the respondents' involvement in and contribution to the survey data community, their data sharing behaviour was surveyed. To be able to do so, respondents were first presented with a filter question on whether they had ever conducted a survey and produced survey data on their own or together with other people. Almost three-quarters (74.42 percent) of respondents confirmed that they had collected data in the past. These 1001 respondents were then asked whether they had shared any of these data. Over half (53.35 percent or 534 individuals) confirmed that they had shared their data.

Since there may be other ways of contributing to the survey data community than sharing data, respondents were additionally presented with a list of 9 possible contributions (plus 1

"other" category that was not included in the calculation of the scale) and they were asked, which of these contributions they had made in the past (Figure 27).

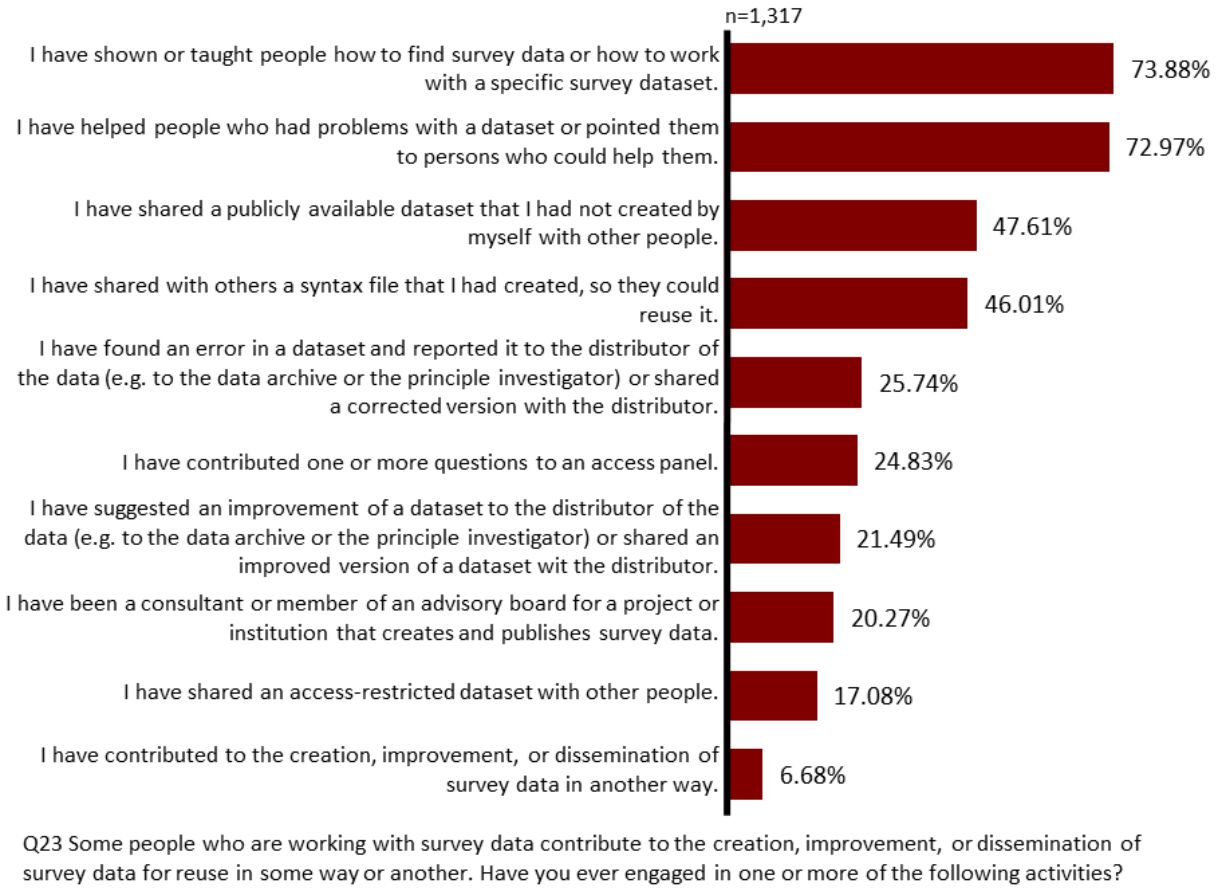


Figure 27 Contributions made to the survey data community

Of the 1,317 valid cases, more than 70 percent each indicated that they had shown or taught people how to find survey data or how to work with a specific dataset (73.88%) or helped people who had problems with a dataset or pointed them to someone who could help them (72.97%). Moreover, almost 50 percent each indicated that they had shared a publicly available dataset (47.61%) or shared a syntax file that they had created (46.01%). The other five possible contributions (reported or corrected errors in datasets; contributed questions to access panels; suggested or made improvements on datasets; consultant or advisory board work; shared access-restricted data) each were mentioned by between 17.08% and 25.74% of cases.

Together with the values from Q21 "Have you ever shared data …" the values of the nine items (excluding the "other" category) from Q23 on contributions to the community comprise a scale of community involvement with values ranging from 0 to 10. The distribution of the community involvement scale is depicted in Figure 28.

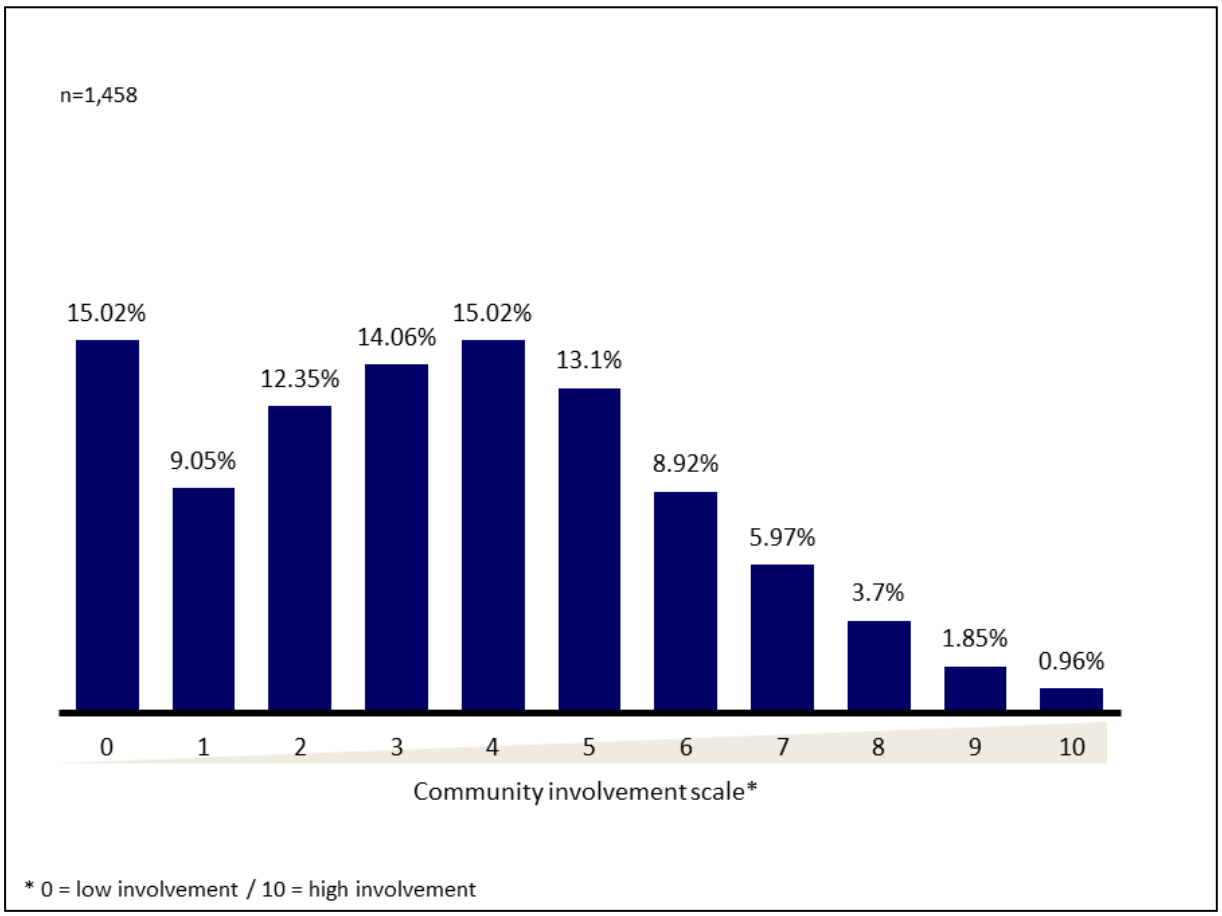


Figure 28 Distribution of community involvement scale

6. Analyses and Results

6.1 The Data Seeking Hypotheses

6.1.1 Hypothesis 1a: Information Seeking through Personal Contact

Hypothesis 1a states: When looking for data, information seeking through personal contact is used more often than impersonal ways of information seeking.

To enquire the respondents' data seeking practices, two separate questions were administered to find out how they had learned about data that they already knew, and where they would look for new data.

To do so, respondents were first asked about their knowledge of a selection of survey programmes (Q7/8). This question was intended as a dimension of experience (see above), but it also served as a prelude to the next question about sources of known data. Participants were asked, where they knew the survey programmes from. The most frequently mentioned source was journal articles (941 mentions, 73.75% of valid cases), followed by colleagues or friends (794 mentions, 62.23% of valid cases) and teachers/professors or supervisors (693 mentions, 54.31% of valid cases) (Figure 29).

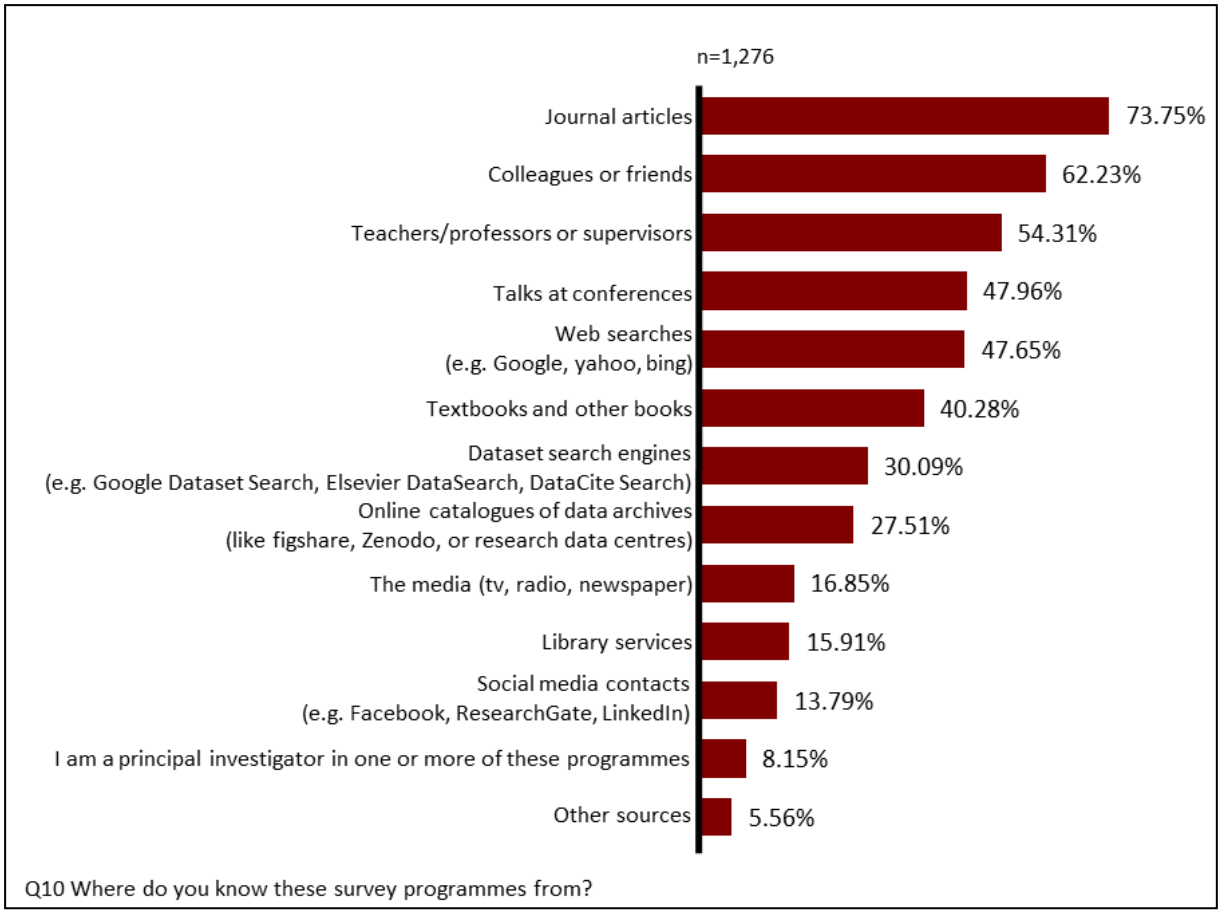


Figure 29 Sources of known surveys

These numbers support the expectation that data users use personal contacts for information seeking, but they fall short of supporting the hypothesis that personal interactions are used more often than impersonal practices of seeking.

The following question was designed to find out more about specific practices when trying to find new data. First, all respondents were asked whether they had searched for survey

data that they could use for their work or for their studies in the past two years (Q11). Over 88 percent (88.07% or 1.284 respondents) reported that they had searched for data in the last two years, 11.93 percent (174 respondents) indicated that they had not. Only those who had done searches were presented with a question on finding data that was very similar to the question on sources of known survey programmes. This one asked respondents which sources they use to find suitable data (Q15).

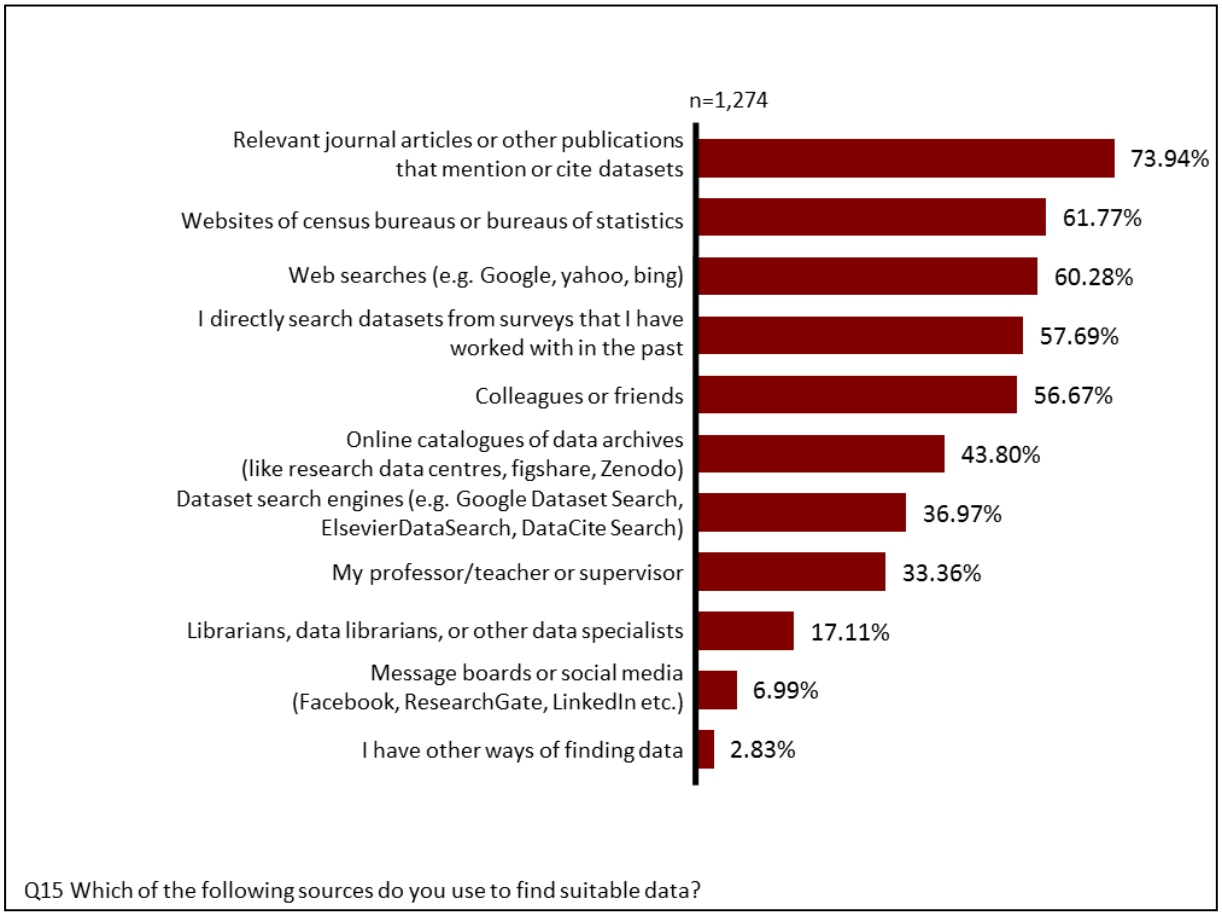


Figure 30 Sources used to find data

Similar to (Q10) that was intended to find out where respondents had learned about known surveys, the most frequently mentioned source of data seeking was again journal articles (942 mentions, 73.94% of valid cases) (Figure 30). Remarkably though, the next most frequent mentions are not personal contacts (colleagues/friends, professors/supervisors) like in the known surveys question; with regard to actively seeking data, searching websites, online catalogues, dataset search engines, or the web in general seem to be more important

than personal contacts. It stands out here that asking colleagues/friends or professors/supervisors is less prominent when actively searching for data (56.67 percent of respondents for colleagues/friends and 33.36 percent for professors/supervisors) compared to the importance of these personal contacts with regard to known surveys (62.23 percent for colleagues/friends and 54.31 percent for professors/supervisors). However, 57.69 percent of respondents indicated that they directly search datasets that they have worked with in the past, which they should know from personal contacts as suggested by the answers to the question on sources of known surveys (Figure 29). Overall this means that using personal contacts is a particularly successful strategy for finding data. Additionally, if people are looking for new data, they also search the web and online catalogues to a substantial extent.

6.1.2 Hypothesis 1b: Personal and Impersonal Ways of Information Seeking by Experience

Hypothesis 1b states: Ways of information seeking (personal or impersonal) differ with experience.

To analyse this hypothesis, correlations were calculated between both questions that enquire practices of data seeking and the experience index. The formation of the experience index is described in detail subchapter D.0. The resulting experience index ranges from 0 (no experience) to 12 (high experience). The distribution of the experience index is depicted in Figure 26.

The correlations of the experience index with the sources of known data that were surveyed with Q10 (Figure 29) are presented in Table 15. The analysis shows that there is a moderate positive correlation between experience and knowing datasets from journal articles (Pearson's $r = 0.3419$), from colleagues or friends ($r = 0.3119$), and from talks at conferences ($r = 0.3072$). Knowing a survey because of one's own role as a principal investigator is weakly positively correlated with experience ($r = 0.1755$) as is knowing surveys from online data catalogues ($r = 0.1562$). Very weak positive correlations with experience can still be found for knowing surveys from dataset search engines ($r = 0.0707$) and from books ($r = 0.0624$). This means that with growing experience, these sources of known data become more relevant. Very weak to weak correlations are found for knowing surveys from professors or supervisors ($r = -0.0556$) and knowing surveys from the media ($r = -0.1160$), which means that these sources tend to be more important for people with less experience.

Table 15 Correlations of sources of known data with experience index

Correlations of sources of known data with experience index	Correlations
Journal articles	0.3419***
Colleagues or friends	0.3119***
Talks at conferences	0.3072***
I am a principal investigator in one or more of these programmes	0.1755***
Online catalogues of data archives	0.1562***
Dataset search engines	0.0707*
Textbooks and other books	0.0624*
Social media contacts (e.g. Facebook, Researchgate, LinkedIn)	0.0512
Web searches (e.g. Google, yahoo, bing)	0.0343
Library services	-0.0257
Teachers/professors or supervisors	-0.0556*
The media	-0.1160***
	Observations 1,276

* p < 0.05, ** p < 0.01, *** p < 0.001

Q10 Where do you know these survey programmes from?

The correlations between experience and knowing surveys from social media contacts, from web searches, or through library services are not significant in the present sample. This means that no clear relationship between these sources of known data and experience can be stated here.

With regard to finding data, the situation looks slightly different. This can be shown with a correlation analysis of the experience index with the sources used to find data (Table 16).

Table 16 Correlations of sources used to find data with experience index

Correlations of sources used to find data with experience index	Correlations
I directly search datasets from surveys that I have worked with in the past.	0.2045***
Colleagues or friends.	0.1988***
Online catalogues of data archives (like research data centres, figshare, Zenodo).	0.1225***
Relevant journal articles or other publications that mention or cite datasets.	0.1026***
Websites of census bureaus or bureaus of statistics.	0.0601*
Message boards or social media (Facebook, ResearchGate, LinkedIn etc.)	0.0046
Dataset search engines (e.g. Google Dataset Search, ElsevierDataSearch, DataCite Search)	-0.0057
Web searches (e.g. Google, yahoo, bing)	-0.0275
Librarians, data librarians, or other data specialists.	-0.0741**
My teacher/professor or supervisor.	-0.1830***
Observations 1,274	
* p < 0.05, ** p < 0.01, *** p < 0.001	

Q15 Which of the following sources do you use to find suitable data?

As has been shown above, directly searching datasets from surveys that they have worked with in the past is a way to find data for 57.69 percent of respondents. The correlation analysis shows that this practice is weakly associated with experience (r = 0.2045). The same is true for asking colleagues or friends (r = 0.1988). Interestingly, the practice of searching journal articles is not as strongly correlated with experience (r = 0.1026) as it is in the case of known data (r = 0.3419). This means that, while actively searching journal articles for suitable data is the most named seeking practice, it is less dependent on experience than knowing data from journal articles. The reason behind this difference may be that in academics, the practice of reading journals grows more important with experience. As a

result, experienced researchers become aware of surveys more often through journal articles without necessarily searching them to find data. Using journal articles to find data is only slightly more important at higher levels of experience. The same is true for searching websites of census bureaus or bureaus of statistics, a practice that is very weakly correlated with experience (r = 0.0601). A very weak negative correlation with experience can be seen for the practice of asking librarians, data librarians, or other data specialists when looking for suitable data (r = -0.0741). The negative correlation with the strategy of asking professors or supervisors for help is more clear (r = -0.1830) and even stronger that for the case of known data. The correlations between experience and using message boards or social media, dataset search engines, or web search engines to find data are not significant in the present sample. This means that no clear relationship between these strategies of data seeking and experience can be stated here.

In sum, correlating sources of known and suitable data with experience shows noticeable differences in the importance of various sources with regard to experience. Not surprisingly, journal articles are more important with growing experience, even though the effect is not as strong for active searches for suitable data as it is for sources of known data. The same is true for colleagues or friends, whose importance as a source of data grows with experience. Not surprisingly, professors and supervisors are more important sources of known or suitable data at lower levels of experience. Interestingly, for the case of finding suitable data, there is also a correlation between lower experience and asking librarians, data librarians, or other data specialists. All in all, personal contacts are important at every level of experience, even though colleagues and friends are more important with growing experience while asking professors, supervisors, or data librarians becomes less important.

6.2 The Experience Hypotheses

In line with the experience hypotheses, it was expected that with growing experience, data users have more ambitious goals, advanced requirements and more specific problems. To answer to these hypotheses, the experience index was correlated with goals (purposes), requirements, and problems.

6.2.1 Hypothesis 2a: Goals and Experience

Hypothesis 2a states: Experience is positively correlated with having ambitious goals.

The goals that users of survey data try to achieve were operationalized as purpose of data use in the past two years. The purposes chosen for this measurement point to ambitious goals such as great academic success (via scientific publication) or basic goals such as developing skills of data use or analysis (via learning).

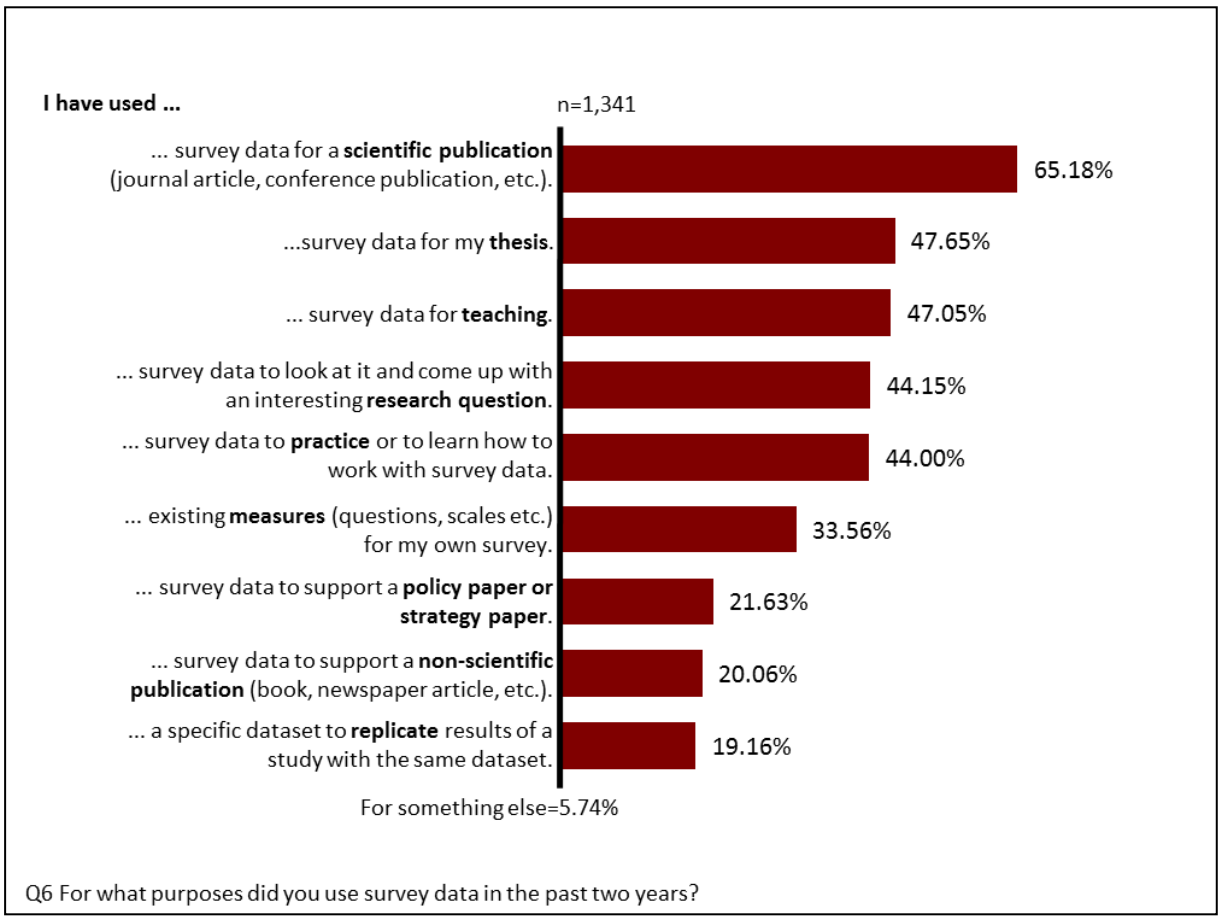


Figure 31 Purpose of data use

Out of those who have used survey data, 1,341 answered the next question on the purposes that they had used survey data for in the past two years. Nearly two-thirds (65.18%) of these indicated that they had used survey data for scientific publications (Figure 31). The next most mentioned purposes are use of data for theses (47.65%), for teaching (47.05%), for coming up with research questions (44.15%), and for practice (44.00%). Use of data for non-scientific publications (20.06%) and to replicate results (19.16%) are the least mentioned purposes. It was expected that more experienced data users had more ambitious goals (or purposes) than less experienced users. The surveyed items had been created to represent

different levels of ambition as described in D.0 and depicted in Table 17. As already explained in D.0, the allocation to the different levels of ambition should be read as a rough approximation rather than an exclusive attribution.

Table 17 Purposes according to levels of ambition

Purpose	Level of ambition
Use of data for scientific publication	High
Use of data to replicate results	High
Use of data for teaching	High
Use of data to come up with research question	Medium
Use of data for thesis	Medium
Use of existing measures	Medium
Use of data for practice	Low
Use of data for policy or strategy paper	Low
Use of data for non-scientific publication	Low

Table 18 shows the pairwise correlations between each of the given purposes and the level of experience as measured with the experience index. All purposes in this table correlate significantly with experience except the purpose of having used data for a thesis. In line with the hypothesis that more experienced data users have more ambitious goals, using data for a scientific publication is strongly correlated with experience (r = 0.5400). Also, the rather ambitious purposes of reusing measures for own data collection and for replication of results are positively correlated with experience (r = 0.2618 resp. r = 0.1647). The positive correlation of using data for teaching with experience stands out (r = 0.3737). The high correlation can be seen because teaching usually comes with the career path towards professorship, which indeed is an ambitious goal. Other purposes that are positively correlated with experience, albeit more weakly, are using survey data for a non-scientific publication (r = 0.1425), to come up with a research question (r = 0.1127), or to support a policy paper or strategy paper (r = 0.0913). The weak negative correlation between experience and the purpose of practicing or learning how to work with survey data (r = -0.1110) also supports the hypothesis.

Table 18 Correlations of purposes of data use in the past two years with experience index

Correlations of purposes of data use in the past two years with experience index	Correlations
Use of data for **scientific publication**	0.5400***
Use of data for **teaching**	0.3737***
Use of existing **measures**	0.2618***
Use of data to **replicate** results	0.1647***
Use of data for **non-scientific publication**	0.1425***
Use of data to come up with **research question**	0.1127***
Use of data for **policy or strategy paper**	0.0913***
Use of data for **thesis**	0.0133
Use of data for **practice**	-0.1110***
Observations 1,341	
* p < 0.05, ** p < 0.01, *** p < 0.001	

Q6 For what purposes did you use survey data in the past two years?

The only purpose that is not significantly correlated with experience in the present sample is the use of data for a thesis. This means that no clear relationship between this purpose and experience can be stated here. A possible reason is that writing a thesis is not associated with one single career stage but occurs during undergraduate studies, graduate studies and PhD studies.

6.2.2 Hypothesis 2b: Requirements and Experience

Hypothesis 2b states: Experience is positively correlated with having more advanced requirements for data.

Respondents were asked to rate requirements when searching for data on a scale from 1 (=not important at all) to 5 (=very important). The analysis of the top two ratings (values 4 and 5) reveals that high data quality, a good fit with the research question, availability free

of charge as well as good documentation and sufficient descriptive information are the most important requirements (Figure 32).

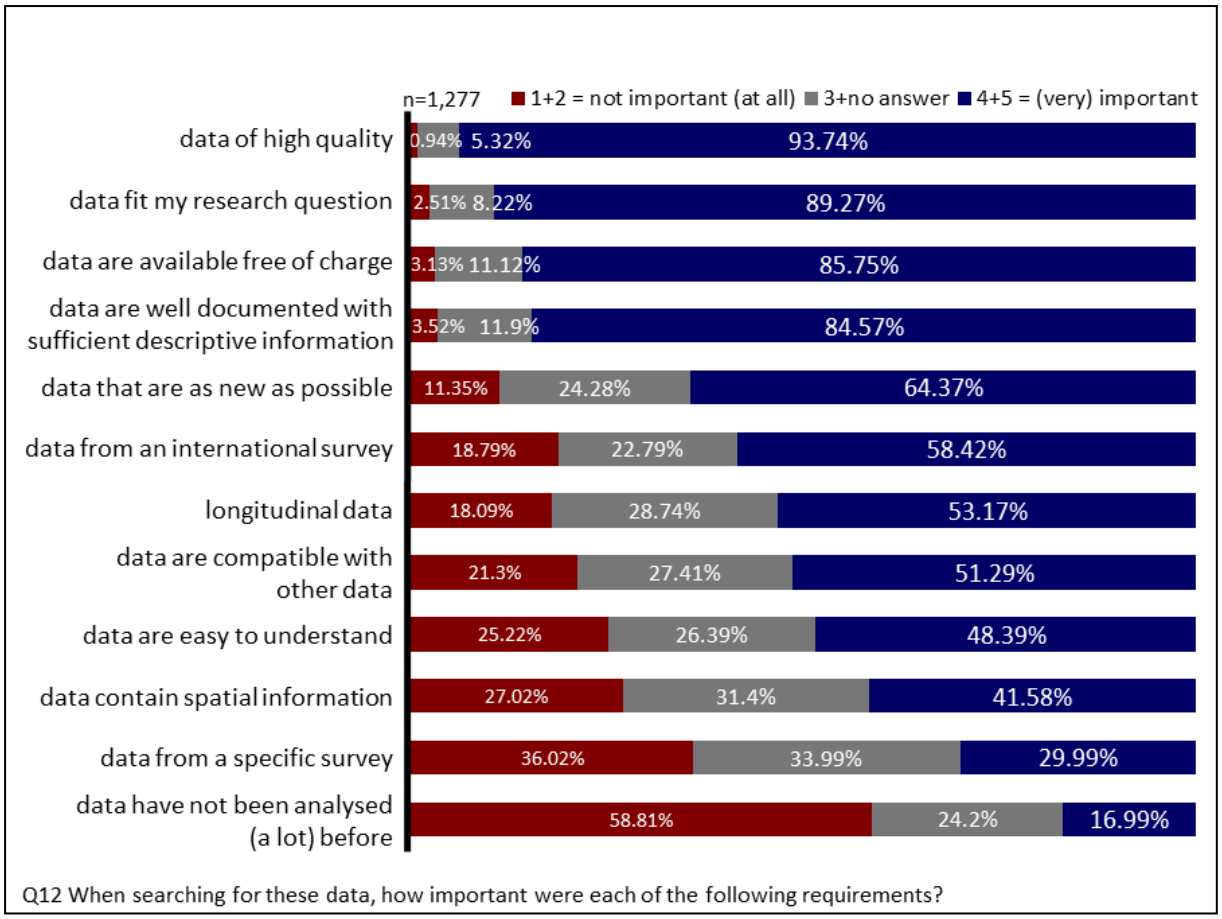


Figure 32 Important requirements when searching for data

Each of these four requirements was indicated as important or very important by more than 80 percent of respondents. The runners up to these criteria are 20 percent or more behind, starting with the requirement of data being as new as possible and ending with the requirement that data should not have been analysed (a lot) before, which was deemed important or very important by only 16.99 percent of valid cases.

It was expected that differing experience is associated with different requirements when searching survey data. Table 19 shows the pairwise correlations of the surveyed requirements with experience.

Table 19 Correlations of requirements when looking for data in the past two years with experience index

Correlations of requirements when looking for data in the past two years with experience index	Correlations	*Observations*
The data should **fit** my research question.	0.1168***	*1,255*
The data should be of high **quality**.	0.1089***	*1,251*
The data should come from an **international** survey, because I wanted to make comparisons between countries.	0.1068***	*1,258*
The data should be well **documented** with sufficient descriptive information.	0.0903**	*1,250*
The data should come from a **longitudinal** survey, because I wanted to study change over time.	0.0259	*1,248*
The data should be available **free of charge**.	0.0087	*1,255*
The data should come from a **specific survey**.	0.0003	*1,232*
The data should be **compatible** with other data that I already had.	-0.0006	*1,241*
The data should **not have been analysed** (a lot) before.	-0.0054	*1,236*
The data should be as **new** as possible.	-0.0894**	*1,251*
The data should contain **spatial** information.	-0.1150***	*1,231*
The data should be **easy** to understand (e.g., results, tables, or simple statistics).	-0.2633***	*1,241*

* p < 0.05, ** p < 0.01, *** p < 0.001

Q12 When searching for these data, how important were each of the following requirements? Please indicate importance on a scale from 1 (not important at all) to 5 (very important).

A very weak to weak positive correlation can be seen for the requirements of data being well documented (r = 0.0903), data coming from international surveys (r = 0.1068), high data quality (r = 0.1089) and that data fit the research question (r = 0.1168). There is weak to moderate negative correlation with the requirements that data should be as new as possible (r = -0.0894), that data should contain spatial information (r = -0.1150) and that data be easy to understand (r = -0.2633). With regard to the hypothesis that differing experience is associated with different requirements, it can be seen that lower experience is indeed

associated with other requirements than higher experience. The strongest effect can be seen for the requirement of data being easy to understand, the other correlations are weaker but significant. Some surveyed requirements do not correlate significantly with experience. This applies to the requirements of data being free of charge, data coming from a specific survey, data from longitudinal surveys, data that are compatible with other data, and data that have not been analysed a lot before.

To further illustrate the relationship between the requirements and experience, t-tests for a difference in mean between groups were calculated for all significantly correlating requirements for two groups of people: a group 1 of 667 respondents who score below average on the experience index (mean = 8.94513) and a group 2 of 791 who score above average on the experience index (Table 20).

Table 20 t-test of requirements for groups of people with experience below or above average

t-test of requirements for groups of people with experience below or above average	**Mean of group 1** (with experience **below** average)	**Mean of group 2** (with experience **above** average)	variance ratio (p)	t-test (p)
The data should **fit** my research question.	4.484112	**4.633333**	0.0013	0.0006
The data should be of high **quality**.	4.657895	**4.744089**	0.0006	0.0115
The data should come from an **international** survey, because I wanted to make comparisons between countries.	3.492593	**3.750696**	0.0993	0.0003
The data should be well **documented** with sufficient descriptive information.	4.295880	**4.432961**	0.0054	0.0049
The data should be as **new** as possible.	**3.895522**	3.758042	0.9200	0.0261
The data should contain **spatial** information.	**3.362101**	3.124642	0.2878	0.0007
The data should be **easy** to understand (e.g., results, tables, or simple statistics).	**3.763158**	3.102962	0.0054	0.0000
Observations	*667*	*791*		

Q12 When searching for these data, how important were each of the following requirements? Please indicate importance on a scale from 1 (not important at all) to 5 (very important).

The t-test confirms the results of the correlation analysis and reveals to what extent people with experience below and above average differ in their requirements when searching for data. The difference is clearest for the requirement of data being easy to understand. Rather unsurprisingly, this requirement is more important for respondents with less experience. A more unexpected finding is that the group with less experience also scores higher on the requirement that data should be as new as possible. The interviewees in the qualitative part of the study had indicated that this was a particular important requirement for experienced researchers. Even more unexpectedly, they also score higher on the requirement that data should contain spatial information. A possible explanation for this is that the answer was phrased ambiguously. What was meant by this answer is that data should contain small-scale spatial information. This kind of data is in fact rarely included in survey data, because it increases the risk of making respondents identifiable. Surveys that include such data are oftentimes only shared in anonymized versions or restrictively provided to researchers in safe rooms. Researchers who work with this kind of data are usually more experienced than others, which is at odds with what is found here. The reason for this may be that respondents understood the answer to mean that data should contain very general spatial or geographical information, such as country codes. Another explanation might be that there are more young researchers who are interested in spatial data, which would point to evidence for a generational shift with younger researchers moving towards using other kinds of data (spatial data, digital behavioural data, etc.).

Looking at the second group, the respondents with experience above average, an interesting finding stands out: more experienced data users rated the requirement that data fit the research question higher than respondents with less experience. A reason for rating this requirement lower may be that people with less experience don't deem lack of fit problematic, because they are more inclined to adjust their research question if they experience problems finding data. This can be shown by calculating the correlation between the problem solving strategy of "adjusting the research question" (Q17) and the experience index. There is indeed a very weak negative relationship between this problem solving strategy and the experience index (r = -0.0787, p = 0.0082). The other requirements that respondents with experience above average have rated higher are: that data be well documented; that data be of high quality; and that they come from international surveys.

The higher rating makes sense for all these requirements. The strongest difference can be seen for the requirement of international data. This is not surprising, because datasets from international surveys tend to be very complex and thus can only be handled by researchers with sufficient experience. The higher score on the quality requirement also makes sense, because more experienced researchers need to work with high quality data if they want to publish their research in renowned journals. Finally, the requirement that data be well documented also corresponds to higher experience. If we assume that more experienced researchers use more complex datasets (from international, longitudinal, or panel surveys), they also need more detailed and reliable documentation for being able to analyse these data. Furthermore, more experienced users are more likely to have had experiences with bad documentation. Therefore, they should be more aware of this problem.

6.2.3 Hypothesis 2c: Problems and Experience

Hypothesis 2c states: Experience is positively correlated with having more specific problems with data.

Respondents were asked what problems they had encountered when finding or accessing survey data (Q16). From a list of possible problems, they were asked to select a maximum of five main problems that they had encountered (Figure 33). The five most frequently indicated problems turned out to be: respondents couldn't find data on their topic of interest (50.87%); description or information on the data was insufficient (40.08%); respondents were denied access to data for legal or other reasons (39.61%); respondents couldn't find data on their population of interest (37.09%); and the data they found were too old (35.75%). Among the three least mentioned problems are lack of knowledge to open or read the dataset (10.87%) and lack of knowledge to understand the content of the dataset (10.16%). Given that these are problems that should correlate with lack of experience, it is no surprise that they score low in the present sample that contains many experienced respondents (see subchapter "D.4.2 Background: Education and Survey Data Literacy").

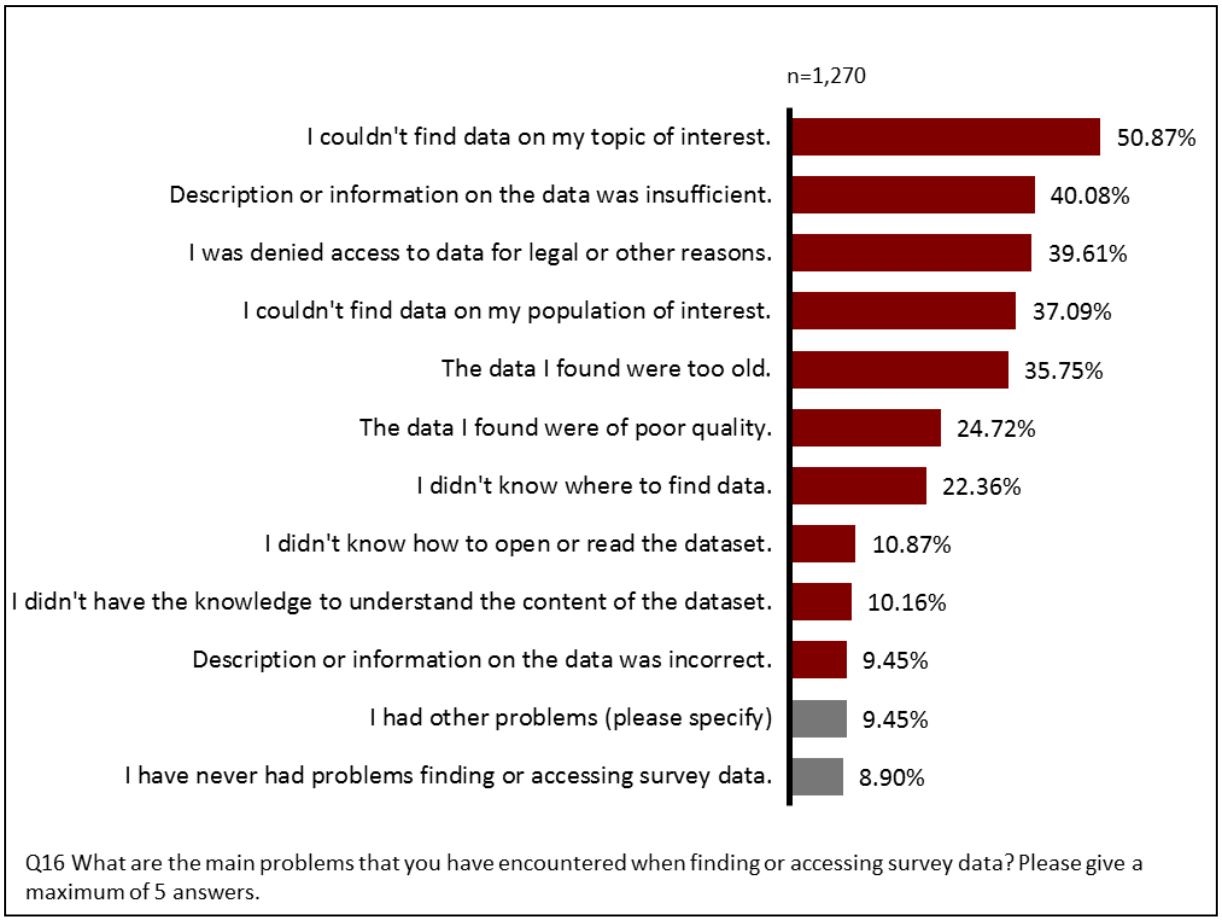


Figure 33 Main problems encountered when finding or accessing survey data

To test the hypothesis, pairwise correlations between each problem and the experience index were calculated. The hypothesis refers to the specificity of problems as it was defined in 0. The assumed order of problems from very general to very specific is again depicted in Table 21.

Table 21 Surveyed problems in ascending order of specificity

I didn’t know where to find data.	Very general
I didn’t know how to open or read the dataset.	
I didn’t have the knowledge to understand the content of the dataset.	
I couldn’t find data on my topic of interest.	
I couldn't find data on my population of interest.	
The data I found were too old.	
The data I found were of poor quality.	
Description or information on the data was insufficient.	
Description or information on the data was incorrect.	
I was denied access to data for legal or other reasons.	Very specific

With regard to the hypothesis, a positive correlation between experience and problem specificity was expected. Table 22 shows the pairwise correlations between the experience index and each variable that indicates a problem with regard to finding or accessing survey data.

Table 22 Correlation of problems when finding or accessing survey data with experience

Correlation of problems when finding or accessing survey data with experience	Correlations
I didn't have the knowledge to **understand** the content of the dataset.	-0.2227***
I didn't know how to **open or read** the dataset.	-0.1587***
I didn't know **where** to find data.	-0.1512***
The data I found were too **old**.	-0.0881**
I was denied **access** to data for legal or other reasons.	0.0254
Description or information on the data was **insufficient**.	0.0376
Description or information on the data was **incorrect**.	0.0543
I couldn't find data on my **population** of interest.	0.0549
The data I found were of poor **quality**.	0.0745**
I couldn't find data on my **topic** of interest.	0.0785**

Observations 1,270
* p < 0.05, ** p < 0.01, *** p < 0.001

Q16 What are the main problems that you have encountered when finding or accessing survey data? Please give a maximum of 5 answers.

Apparently, less experienced survey data users have very basic problems of where to find data, how to open or read datasets, and lack of knowledge to understand data. All three variables show a weak to moderate negative correlation with the experience index (Pearson's *r*: -0.1527, -0.1537, and -0.2153). This analysis supports the assumption made in the model that inexperienced data users tend to have very general, basic problems. The

problem of finding data that were too old also has a very weak negative correlation with experience (*r* = -0.0867), which had not been expected initially.

Looking at positive correlations with experience, the case could be made that more experienced users have the more specific problem of poor data quality (*r* = 0.0716). The problem of finding data on a topic is not very specific, but there is a very weak positive correlation with experience as well (*r* = 0.0664). It is possible that the items chosen to represent specific problems are not the most relevant issues for the population and that other items would have shown the expected effect. For instance, after evaluating the open answers given to this question ("other problems"), a better item could have dealt with problems of harmonization or compatibility of data (examples of answers in this category: "same questions were not repeated on different years"; "Need to manually collate multiyear data across different files (Eurobarometer data)"; "lack of harmonization").

Some problems do not correlate significantly with experience in the present sample. This applies to the problems of not finding data on the population of interest, insufficient documentation or information on the data, incorrect documentation or information on the data, and being denied access to data for legal or other reasons.

6.3 The Community Involvement Hypothesis

Hypothesis 3 holds that Experience is positively correlated with community involvement.

People with much experience in data use and analysis were expected to be more involved in the community than people with less experience. The analysis of this hypothesis again draws on the experience index and on the community involvement scale that has been introduced above (D.0).

This scale combines measurements of two questions. First, respondents were asked whether they had ever shared data that they had collected. This question was administered to those 1,001 respondents (74.42 percent of the whole sample) who had indicated that they had collected data in the past. Of these respondents, over half (53.35 percent or 534 individuals) confirmed that they had shared their data. Since there may be other ways of contributing to the survey data community than sharing data, respondents were additionally presented with a list of 9 possible contributions (plus 1 "other" category that was not included in the calculation of the scale) and they were asked, which of these contributions they had made in

the past (Figure 27). Both measures were combined to a scale of community involvement that ranges from 0 (no involvement) to 10 (high involvement). The distribution of the community involvement scale is depicted in Figure 28.

A simple linear regression of these two measures was calculated to analyse the relationship between community involvement and experience. A significant regression equation was found ($F(1, 1456) = 716.12$, $p < 0.000$), with an R^2 of 0.3297. Participants' community involvement increased 0.5482703 with each point on the experience index (Figure 34). The R^2 indicates that experience accounts for increase in community involvement to the extent of 32.97 percent.

It can be concluded from this analysis that there is indeed a positive relationship between experience and community involvement as stated in hypothesis 3.

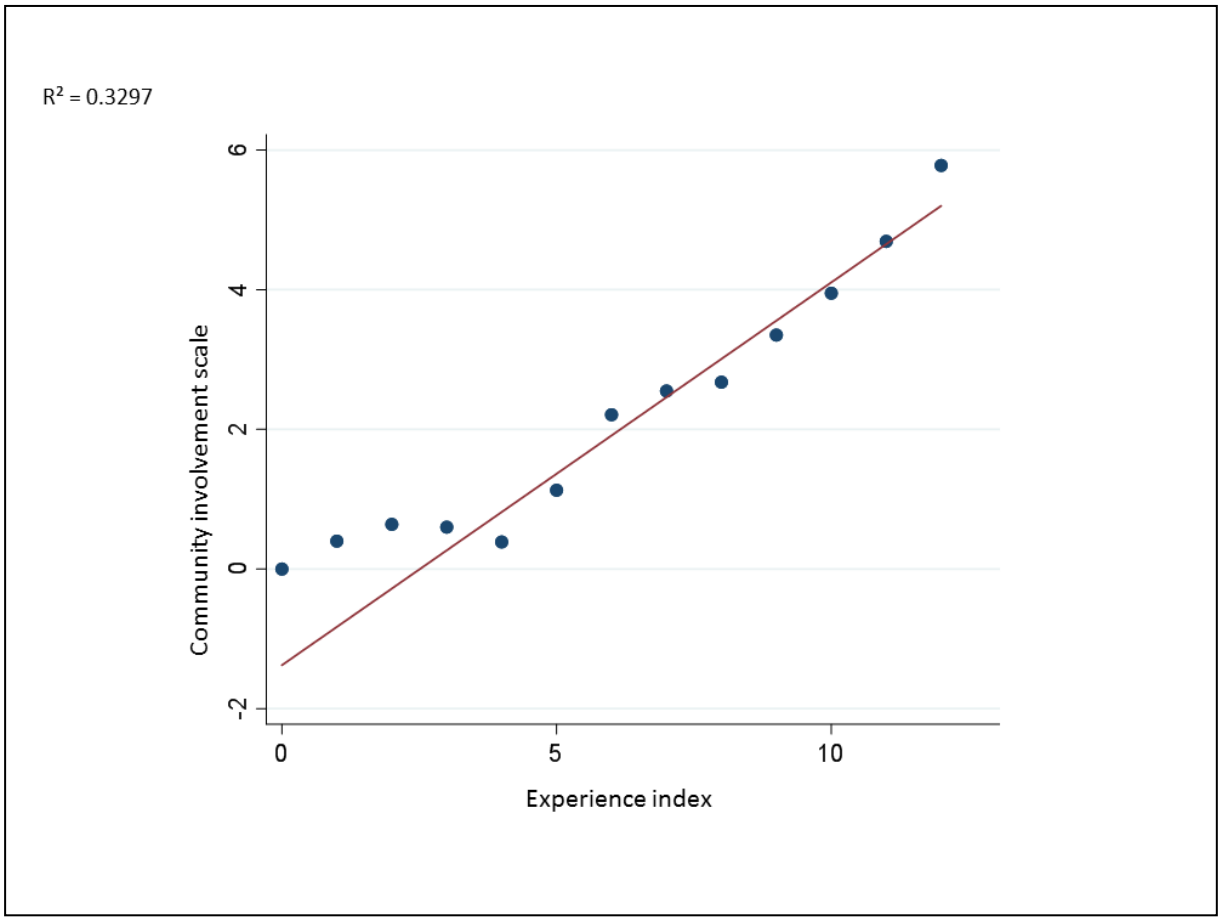


Figure 34 Regression of community involvement index and experience index (scatter plot with regression line)

6.4 The Problem Solving Hypothesis

Hypothesis 4 holds that Community involvement is positively correlated with problem solving strategies that require personal interactions.

This hypothesis was investigated in two steps. The first step was to find out whether community involvement (measured by the community involvement scale) influences the choice or valuing of specific problem solving strategies. In a second step, the influence of community involvement on success in problem solving was analysed.

To investigate the respondents' choice of problem solving strategies, they were asked how they dealt with problems of finding or accessing survey data (Q17). The respondents were presented with a list of strategies and were asked to rate the importance of these strategies on a scale from 1 (=not important at all) to 5 (=very important).

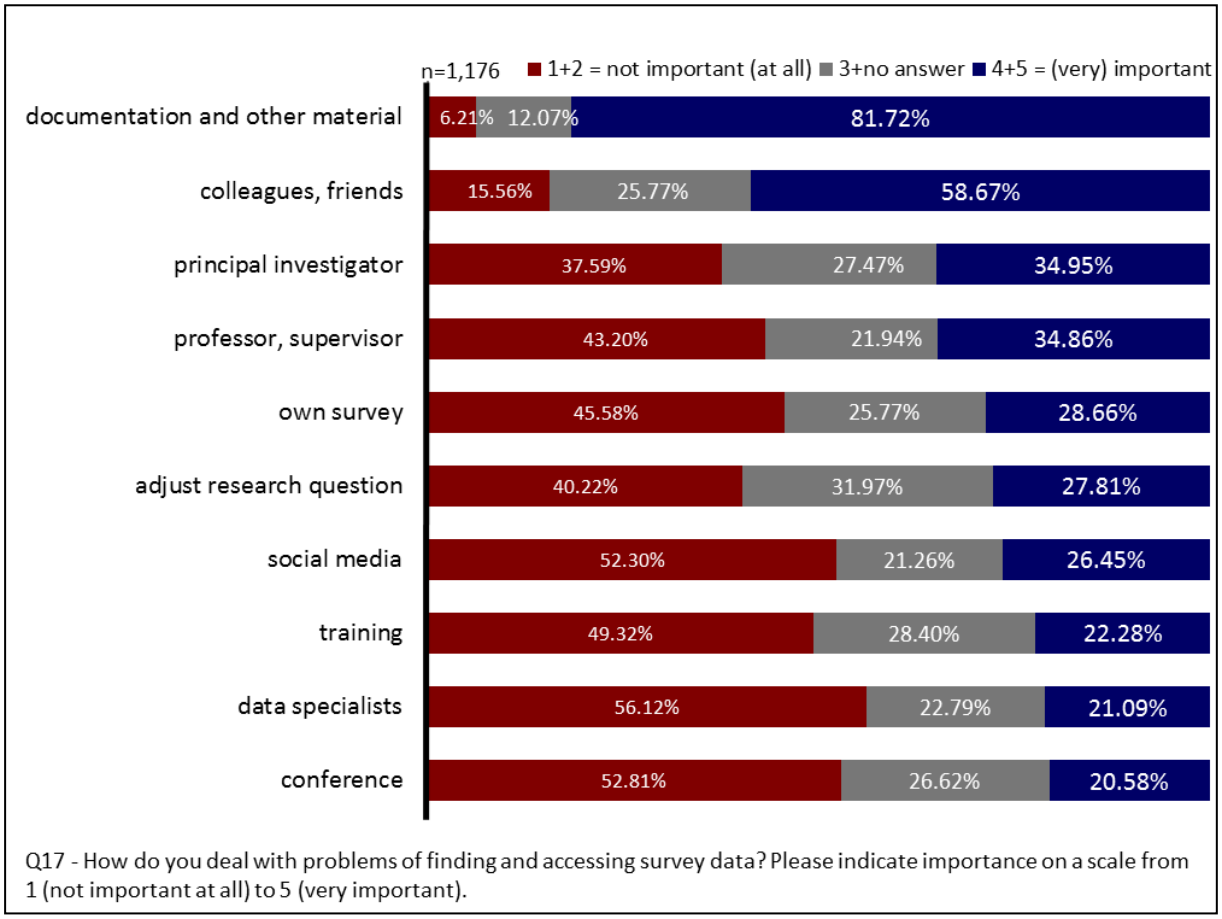


Figure 35 Important strategies of dealing with problems of finding and accessing survey data

In an evaluation of the top two ratings (values 4 and 5), the most important strategies are consulting documentation and other information material (81.72% of valid cases) as well as asking colleagues or friends for help (58.67% of valid cases) (Figure 35). The least important strategies are participation in training measures (22.28% of valid cases), consulting data librarians or other data specialists (21.09% of valid cases), and visiting a conference that deals with the survey (20.58% of valid cases).

In order to test the assumption that community involvement is correlated with valuing specific problem solving strategies, correlations between community involvement and problem solving strategies were calculated. In the analysis, only those strategies from Q17 were included that require personal interactions, because only these can be seen as open to influence by community involvement (see definition of problem solving in D.0). The relevant strategies are:

- Finding help online (social media or message boards)
- Participating in training/ a workshop that deals with this problem
- Visiting a conference/ event that deals with the data/ dataset
- Asking professors or supervisors for help
- Asking colleagues or friends for help
- Asking data librarians or other data specialists for help
- Asking the person who collected the data (principal investigator) for help

To find out whether the value of these personal interactions for solving problems is correlated with community involvement, a correlation analysis between each of these strategies and the community involvement scale was calculated (Table 23). The analysis shows that there is a positive correlation between community involvement and the strategies of asking colleagues or friends for help (r = 0.0636, very weak correlation) as well as asking the person who collected the data (principal investigator) for help (r = 0.1444, weak correlation). This means that people with stronger community involvement tend to value these two strategies more. Community involvement is negatively correlated with the strategy of asking a professor or supervisor for help (r = -0.2292). This is not surprising since it has already been established that community involvement grows with experience and more experienced people tend not to ask professors, teachers or supervisors for assistance.

Table 23 Correlation of problem solving with community involvement

Correlation of problem solving with community involvement	Correlations	*Observations*
I ask the person who collected the data for help.	**0.1444*** **	*1,128*
I ask colleagues or friends for help.	**0.0636***	*1,135*
I ask my professors/teachers or supervisors for help.	**-0.2292*** **	*1,112*
I try to find help in online message boards or social media (Facebook, ResearchGate, LinkedIn, etc.).	-0.0520	*1,123*
I participate in training or a workshop that deals with this problem.	-0.0319	*1,121*
I ask data librarians or other data specialists for help.	-0.0131	*1,119*
I visit a conference or another event that deals with the survey data that I want to work with.	0.0442	*1,122*

* p < 0.05, ** p < 0.01, *** p < 0.001

Q18 How do you deal with problems of finding and accessing survey data? Please indicate how important the following strategies of problem solving are for you on a scale from 1 (not important at all) to 5 (very important).

Four strategies are not significantly correlated with community involvement: finding help in online message boards or on social media; participating in training or workshops; asking data librarians and other specialists; visiting a conference or other event that deals with the survey data in question. For these strategies, no clear relationship with community involvement could be shown in the current sample.

The findings partially support the hypothesis that community involvement is correlated with problem solving strategies that require personal interactions. To get an even better picture of the value of community involvement for problem-solving, further analyses were made with regard to the specificity of problems as defined earlier (see D.0). It is assumed here that people benefit all the more from community involvement if they have to solve more specific problems. The most specific problems surveyed with question Q16 (Figure 33) were:

- Dealing with insufficient documentation or information on a dataset
- Dealing with incorrect documentation or information on a dataset
- Being denied access to data for legal or other reasons

To determine whether the valuing of specific problem solving strategies helps to resolve specific problems in particular, the correlations were again calculated for this restricted set of problems. It was expected that for people with these problems the already shown rather weak correlation of community involvement and problem solving would be stronger. This would mean that community involvement positively influences problem solving in particular with regard to more specific problems. To determine, how strongly community involvement and problem solving correlate with regard to these problems, correlations were calculated only for those people who had indicated having one or more of these problems. Table 24 shows the correlations of problem solving strategies with community involvement for those people who had indicated a specific problem.

Table 24 Correlations of problem solving strategies with community involvement for respondents with specific problem

Correlations of specific problem solving strategies with community involvement for respondents who have indicated to have specific problems	**Correlations**	*Observations*
Respondent indicated to have had this problem: **Being denied access to data for legal or other reasons**		
I ask the person who collected the data for help.	**0.1548*** **	*478*
Respondent indicated to have had this problem: **Dealing with insufficient documentation or information on a dataset**		
I ask the person who collected the data for help.	**0.1947*** **	*489*

* p < 0.05, ** p < 0.01, *** p < 0.001

Q18 How do you deal with problems of finding and accessing survey data? Please indicate how important the following strategies of problem solving are for you on a scale from 1 (not important at all) to 5 (very important).

The only strategy that showed significant correlations in this analysis was asking the person who collected the data (principal investigator) for help. There was no significant correlation of asking the principal investigator with the problem of incorrect documentation or information on a dataset. For people who have indicated that they had been denied access to data or had to deal with insufficient information on a dataset, community involvement is significantly positively correlated with the strategy of asking the person who collected the data (principal investigator) for help:

- Being denied access to data for legal or other reasons ($r = 0.1548$, $p < 0.001$)
- Dealing with insufficient documentation or information on a dataset ($r = 0.1947$, $p < 0.001$)

In both cases the positive correlation with the community involvement scale is stronger than for people who have not indicated to have had these problems ($r = 0.1444$ for the strategy of asking the person who collected the data, see Table 23). For these problems, the analysis sufficiently confirms the hypothesis that community involvement is positively correlated with problem solving strategies that require personal interactions.

7. Findings

The quantitative study was set up to test the following hypotheses:

(1) The data seeking hypotheses:

(1a) When looking for data, information seeking through personal contact is used more often than impersonal ways of information seeking.

This hypothesis was confirmed with certain qualifications. As it turns out, the most used way to find data is through journal articles. This can be seen in the answers given to two questions. The first question, "Q10 Where do you know these surveys from?", referred to the knowledge of popular surveys that had been presented to the respondents before. The second question, "Q15 Which of the following sources do you use to find suitable data?", was used to find out how respondents were looking for data in general. Remarkably, the extent to which practices of seeking through personal contacts are used is considerably smaller for the second question than for the first. Apart from searching journals, the

practices of searching the web (including websites of bureaus of statistics) as well as directly searching known datasets are used more than personal contacts (colleagues, friends, professors, supervisors). However, searching known datasets again points to the first question, where personal contacts as sources of known data were indicated more often than web searches and other impersonal ways of searching. This suggests that, apart from searching journals, data seeking through personal contacts is more successful than searching the web or other impersonal ways of seeking (for instance, through data catalogues). With regard to the hypothesis it can be concluded that personal contacts are not necessarily used more often, but they are a very important, indispensable part of survey data users' information seeking behaviour.

(1b) Ways of information seeking (personal or impersonal) differ with experience.

With regard to the second data seeking hypothesis, the analyses showed that seeking practices indeed differ along the spectrum of experience. With regard to sources of known data, more experience is positively associated with knowing data from conferences and colleagues and with being a personal investigator of a large survey programme. These three correlations show that more experienced data users indeed become aware of survey data through personal involvement. For the case of actively searching for data, it seems that personal contacts are more important for less experienced users; less experience is associated with asking professors/supervisors or data specialists (e.g., data librarians). More experienced users turn to data from surveys that they have already worked with in the past (of all practices, this one has the strongest positive correlation with experience). The only personal way of finding data that is positively correlated with experience is asking colleagues. However, experienced users learned of the known data that they repeatedly use from their personal contacts.

(2) The experience hypotheses:

(2a) Experience is positively correlated with having ambitious goals.

With regards to goals it could be shown that a high score on the experience index correlates with ambitious goals (scientific publishing, own data collection, replication of results). In

turn, the less ambitious goal of learning how to work with survey data correlates with low experience.

(2b) Experience is positively correlated with having more advanced requirements for data.

A correlation analysis and t-tests confirmed the assumption of different requirements in relation to experience. It could be shown that less experience is associated with the requirements that data be easy to understand and that data be as new as possible. The higher value of the requirement of new data is surprising, but probably best explained by the fact that the methods that more experienced researchers use require other, more specific data features that are just more important than currency. The more specific requirements that higher experience is associated with are that data be of high quality, come from international surveys, and be well documented. Surprisingly at first, more experienced researchers also score higher on the requirement that data fit their research question, which was expected to be a general requirement, regardless of experience. But as a correlation analysis has shown, less experienced respondents are more inclined to adjust their research question if they have problems finding the data that they need. This can be interpreted as evidence that they are more inclined to change their research question, if the only data that they find doesn't fit and thus don't rate this problem very highly.

(2c) Experience is positively correlated with having more specific problems with data.

This hypothesis was confirmed in large parts. Three of the surveyed problems did not show significant correlations with experience, which means that these problems are not clearly associated with experience. It could be shown that low experience is associated with very general, basic problems (where to find data, how to open or read datasets, problems to understand data). It could also be shown that high experience correlates with one of the specific problems that were presented to the respondents (poor data quality). Possibly, the surveyed items did not include the most important problems of very experienced survey data users. An evaluation of the open answers on the problem question revealed that for instance, problems of harmonisation or compatibility of data are common. A follow-up study should consider this problem.

(3) The community involvement hypothesis: Experience is positively correlated with community involvement.

This hypothesis is clearly confirmed by the data. The linear regression that was calculated showed a clear positive relationship between experience and community involvement. Experience accounts for an increase in community involvement to the extent of 33.39 percent. With every point on the experience index (ranging from 2 to 12), community involvement increases by 0.58 points (on the community involvement scale ranging from 0 to 10). It could be shown that this effect is also highly significant ($p < 0.000$). It can be drawn from this analysis that people who are working with survey data and grow more experienced over time also increase their community involvement. The data show that more than 70 percent of all respondents engage in showing or teaching others how to find or work with data. Also more than 70 percent help people who have problems with datasets or point them to someone who can help them. Almost 50 percent of all respondents share syntax files with others. Furthermore, from all respondents who had indicated that they had collected data in the past (1001 respondents), more than 50 percent indicated that they had shared these data (mostly with colleagues or friends). Community involvement as measured with these items seems to be an integral element of a survey data researcher's work life.

(4) The problem solving hypothesis: Community involvement is positively correlated with problem solving strategies that require personal interactions.

This hypothesis was confirmed for specific strategies of problem solving that require personal interaction. A correlation analysis confirmed that, in general, community involvement is positively associated with asking colleagues or friends for help as well as with asking the principal investigator. It could be shown that there is an increase in this effect for respondents with very specific problems. For those respondents who have indicated to have been denied access to data, the correlation of community involvement with the strategy of asking the principal investigator for help is stronger. The same is true for respondents who have indicated having had problems with insufficient documentation or information on data. The first result probably is the most interesting one with regard to the theory of problem solving by community involvement. It supports the assumption of the theory that being able to access data not available to everyone is easier for people who are more involved in the

community, because they are in a better position to directly ask the principal investigators. The second result is interesting as well, because it shows that insufficient documentation is not sufficient reason to refrain from using a specific dataset, if you are in a position to ask the right people.

E. Discussion of Results

This study was intended to create empirical evidence for patterns of data-related information seeking behaviour of survey data users. The general research question was: What are the characteristics of information seeking behaviour with regard to survey data? The study was open to investigate patterns as well as stages that occur in data seeking practices and behaviours. It was especially designed to find out more about factors that influence these practices and behaviours. In particular, individual characteristics, social and situational contexts, purposes, goals and problems of survey data users were investigated.

Due to lack of prior research, the investigation was designed as a mixed methods study, more precisely, an exploratory sequential design was used. First, a qualitative study was carried out to develop a grounded theory of survey data seeking. On these grounds, a second quantitative study was conducted that aimed at exemplifying and testing the theory. The results of both studies are discussed in the following paragraphs, leading up to the presentation of a consolidated model of data seeking behaviour and, most importantly, to recommendations for research data infrastructure design.

1. Research Questions and Answers

The specific research question that was investigated in this study is: What are the characteristics of researchers' information seeking behaviour with regard to survey data? Specifically, the study was aimed at finding out how these characteristics and practices depend on social, interactive and contextual parameters. Guiding questions for the characteristics and the influencing, contextual factors posed in the beginning were:

- What patterns occur in data seeking practices/ behaviours?
- What individual characteristics do survey data users have?
- What are the (social, situational) contexts of survey data users?
- What needs do survey data users have, what goals do they try to reach and what purposes do they pursue?
- What are requirements of survey data users who want to find data for reuse?
- What problems do survey data users encounter when looking for data? How do they solve them?

Both the qualitative and the quantitative study gave answers to these questions.

1.1 Patterns in Data Seeking Practices

In chapter "B. Theoretical Perspective", past research on information seeking behaviour was analysed with regard to applicability for data seeking practices. Characteristics and patterns have been common categories for analysis of information seeking behaviour, for example by David Ellis (Ellis 1989) and others.

At the outset of this study, the review of past research led to the expectation that people who are looking for reusable data practice forward chaining from journal articles. This assumption was confirmed in the qualitative study and the quantitative study revealed that this kind of forward chaining was indeed the most mentioned data seeking practice for both known data and previously unknown datasets. It was also expected that the pattern of involving personal contacts or intermediaries played a major role in finding data. The qualitative study confirmed this assumption in that it suggested that students learned about data from professors and more advanced researchers engaged heavily in contacting peers, for example by attending conferences or contacting them personally. The quantitative study showed that colleagues, friends, professors, or supervisors as well as conferences were indeed very important sources of data. It was also expected from the beginning that survey data reusers to a relevant extent relied on data from studies that they had worked with in the past. The qualitative as well as the quantitative study confirmed this assumption and suggest that the academic upbringing in survey research includes becoming familiar with a certain set of reusable data from more or less popular survey programmes. The initial expectation that people who are looking for data are searching the web for reusable datasets, somewhat regardless of experience or professional status, was confirmed as well.

1.2 Individual Characteristics of Survey Data Users

The individual characteristics of survey data users were studied as independent variables that possibly influence data seeking behaviour. In the qualitative part of the study, the experts had already suggested different practices of data seeking behaviour with regard to individual characteristics, in particular with regard to experience. The experts indicated that student users sometimes presented them with their assignment questions and didn't know where to find appropriate data. Sometimes students even ask data archive staff for interesting topics that they could investigate. The participants of the quantitative study were

sampled from registered users of a German data archive. Via its catalogue, this archive provides access to datasets from approximately 6,000 national and international studies from the social sciences. This sampling method resulted in a sample of respondents whose overall level of experience with survey data analysis and research proficiency is very high. About 50 percent of the sample have a doctoral or higher degree and about 41 percent are university or college professors. More specifically, more than 80% replied that they had used expert methods to analyse survey data, which means that the overall level of survey data literacy in the sample is very high. To find out more about the influence of individual characteristics on survey data seeking, the users' proficiency and experience was taken into account when analysing practices of information seeking, goals, requirements, and problems. With regard to practices of information seeking it could be shown that with growing experience, journal articles as well as colleagues or friends are more important sources when looking for data. Professors and supervisors are more important sources at lower levels of experience. There is also a correlation between lower experience and asking librarians, data librarians, or other data specialists when looking for suitable data.

1.3 Contexts of Survey Data Users

The study was conducted from a social-constructivist perspective, which accounts for the general assumption held that context factors are relevant influencing variables in data seeking behaviour. Theoretical considerations made in the beginning already pointed to the relevance of the social context with regard to the roles of intermediaries in data seeking. The results of the qualitative study support this assumption. The interviewed experts made it clear that personal contacts with peers, professors, supervisors, principal investigators, and data professionals played a key role in data seeking practices. For instance, it was found that students are introduced to popular datasets by their teachers or they ask data professionals for help. It was also found that very involved users may have options to find or access data that are inaccessible to others. These results led to the development of the grounded theory of problem-solving by community involvement, from which the hypotheses for the quantitative study were deducted. Both the theory and the hypotheses are presented in detail in chapter C. In the quantitative study, the anticipated roles of intermediaries and of community involvement were confirmed. First of all it was shown that survey data users increasingly become active parts of data communities with growing experience in survey

research. It was also shown that they use their involvement in these communities to find data and to solve problems. The responses indicate that survey data users value asking colleagues, friends, principal investigators, and professors or supervisors almost above any other problem solving strategy. Only consulting the documentation is important or very important to more respondents. In particular, it could be shown that contacting principal investigators is a strategy of problem solving that is especially important to very involved community members. Given that some data are only accessible through the principal investigators who have collected the data, this finding shows how important community structures are for researchers who want to conduct original research, who are looking for new and interesting findings, or simply want to get published with unique results.

1.4 Needs, Goals and Purposes of Survey Data Users

Needs, goals, and purposes have always been important concepts in theories of information seeking behaviour, as has been explained in chapter "B. Theoretical perspective". At the outset of this study, information seeking behaviour was presented as goal-oriented problem solving, making goals a particular relevant concept for the present study. The qualitative study revealed that the users' goals should be dependent on the users' experience in survey research. People with no experience in survey data use should have very basic goals such as obtaining empirical results on a certain topic. Novice users like undergraduate students would have rather moderate goals such as graduating. On the other end of the spectrum, experts in survey research have ambitious goals such as great academic success or innovative and outstanding findings. To measure the level of ambition in goals, nine purposes of data reuse were identified to represent three levels of ambition for the quantitative study. The analyses confirmed the assumption that more experienced data users have more ambitious goals, because they use data for scientific publication or for replication of results. By contrast, less experience is associated with less ambitious purposes such as use of data for practice.

1.5 Requirements of Survey Data Users

With regard to requirements that survey data users have when they are trying to find reusable data, it was expected that context factors would again play an important role. In particular, it was expected that relevance criteria regarding data quality, topics, and methodology would be influenced by the domain and social context. It was also expected

that more experienced researchers would look for more compatibility and comparability in a dataset. The qualitative study confirmed these assumptions in that experts reported that more experienced researchers especially requested high quality datasets as well as datasets that they could use to apply specific methods. Topical relevance was reported to be important as well, but regardless of experience. The quantitative analyses confirmed that there was a correlation between possible requirements when looking for data and the level of individual experience in survey research. As it turns out, more experience correlates with the requirements of topical fit, high quality and (international) comparability. Interestingly, experience is also positively correlated with the requirement of good documentation. As expected, less experienced users mainly require data that are easy to understand.

1.6 Problems and Problem Solving of Survey Data Users

Two other important concepts in information seeking behaviour that were in the focus of this investigation are problems and problem solving. As explained in chapter "B. Theoretical perspective", problems and problem solving have long been at the core of analysing information seeking behaviour. It was expected here, that the nature of the problems and the strategies of problem solving would be specific with regard to survey data. The problems or barriers that users might experience when looking for data were expected to go beyond missing data on a given subject or population. Problems with data quality, data access, data complexity and data documentation were expected to be of relevance. The qualitative study confirmed these expectations. The experts suggested that there was again a correlation with data users' experience, in that less experienced users would have problems of little complexity and expert users would have very specific problems. For example, users with no experience would lack basic skills of data analysis, while expert researchers had to struggle with legal or ethical barriers. These problems can also be tied back to the goals that users on the spectrum of low to high experience might have. The quantitative analysis confirmed these results. Less experience is associated with basic problems such as lacking knowledge of understanding data or opening or reading a dataset. There is also a negative correlation between not knowing where to find data and experience. People with higher experience tend to have problems with data quality. With regard to the solving of these problems, it was expected from a theoretical point of view, that documentation, intermediaries, and information technology should play important roles. In the qualitative as well as the

quantitative study it could be shown that documentation is one of the most important, if not the most important strategy of solving problems when looking for data. However, the qualitative study suggested that strategies that are related to community involvement should be important as well, in particular for more experienced researchers. The results of the quantitative study show that strategies that involve personal contacts are indeed very important. It seems that in particular, users who have problems with data access or with insufficient documentation benefit from asking principal investigators for help. These are problems that prevail among more experienced researchers. These researchers benefit from their community involvement when trying to solve these problems.

2. Theory and Model of the Information Seeking Behaviour of Survey Data Users

2.1 Development and Testing of the Theory

Based on theoretical assumptions that were drawn from general research in information seeking behaviour and from knowledge about survey research and survey data reuse, qualitative interviews were conducted with people who work in data service. These expert interviews yielded rich and multifaceted data with regard to the research question. In particular, the experts gave valuable approximations of patterns and stages that occur in data seeking practices, of the users' purposes, goals and problems. With regard to the individual characteristics of the users, their experience with data reuse and analysis as well as their seniority as researchers were identified as relevant factors with regard to data seeking.

By use of appropriate methodology, a grounded theory was developed on the grounds of these interview data. The developed “Theory of problem-solving by community involvement” consists of several corner stones. The core concepts of the theory, their relations and the nexus between them were condensed in a diagram that represents the model of problem-solving by community involvement (Figure 36).

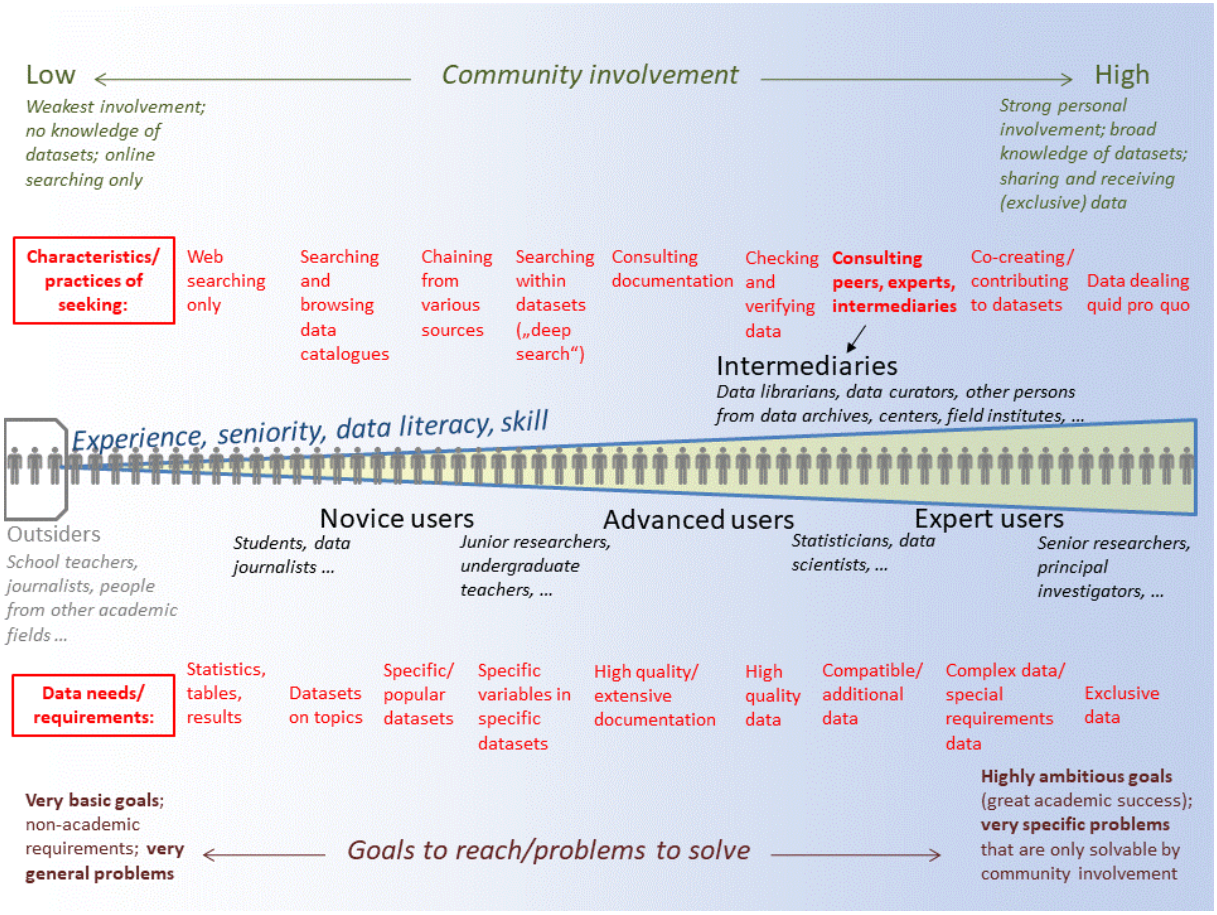


Figure 36 Model of problem-solving by community involvement

According to the model, re-users of survey data are found on a spectrum from outsiders to expert users. Depending on their experience, they have different goals to reach and problems to solve. Their information needs and requirements differ accordingly. Personal interaction with others is a significant factor in goal development, goal achievement and problem resolution for researchers who want to reuse survey data. How users are looking for data in terms of characteristics or practices of seeking is depending on their experience or seniority. Information seeking is manifest in different characteristics or practices and is facilitated through the existence of vital communities surrounding large survey programmes. Survey data communities emerge and persist, because knowledge of them is handed down from senior researchers to junior researchers or shared between peers. With growing experience, seniority, and data literacy, community involvement is increasing. Community involvement facilitates goal-oriented problem solving, and hence information seeking, with regard to survey data. The more specific and delicate the problems are, the less likely it

seems to be that they will be solvable by merely relying on the information provided online or through formal information systems. Being an active community member can improve a researchers' outcomes, because sometimes data are only available through personal contacts.

On these grounds, the following hypotheses for the quantitative study were phrased (Table 25):

Table 25 Hypotheses on data seeking practices and community involvement

(1) The data seeking hypotheses:
(1a) When looking for data, information seeking through personal contact is used more often than impersonal ways of information seeking.
(1b) Ways of information seeking (personal or impersonal) differ with experience.
(2) The experience hypotheses:
(2a) Experience is positively correlated with having ambitious goals.
(2b) Experience is positively correlated with having more advanced requirements for data.
(2c) Experience is positively correlated with having more specific problems with data.
(3) The community involvement hypothesis:
Experience is positively correlated with community involvement.
(4) The problem solving hypothesis:
Community involvement is positively correlated with problem solving strategies that require personal interactions.

The questionnaire for the quantitative part of the study was designed to deliver the data to test these hypotheses. The resulting questionnaire was administered to registered users of a data catalogue that provides access to reusable survey data. With 1,458 completed interviews this survey yielded a comfortable data basis to address the hypotheses.

Overall, the hypotheses were confirmed to varying extent. With regard to part 1a of (1) the data seeking hypotheses, it can be stated that personal contacts are a very important source of known data, and also important sources when looking for new data. However, the most important source for both scenarios (known data and new data) is journal articles. Then again, journal articles are an integral part of scientific communication. So it could even be

acceptable to interpret this source as a community mediated communication, which would be closer to personal than impersonal sources of information. This being said, searching the web as a clearly impersonal source of information is also a very frequently used practice when looking for data. However, it does not seem to be the most successful strategy, given that it ranges behind personal contacts for sources of known data. In subchapter "C.3.2.1 Key Findings and Hypotheses" the assumption had already been made that searching the web (or data catalogues) for data is oftentimes unsuccessful because of a lack of sufficient and standardized survey data documentation. With regard to the second part of the hypothesis (1b) it was found that different levels of experience are indeed associated with different practices of seeking. Personal contacts and involvement play an important role as sources and mechanisms of personal knowledge of the survey landscape (the question of "known data"). Interestingly, more experienced users are more likely to know survey programmes from online data catalogues than less experienced users. With regard to actively seeking data (the question of "how to find data"), personal contacts are important at all levels of experience, but involve different contact persons (professors/supervisors, if less experienced; and colleagues/friends, if more experienced). Asking data librarians or other data specialists for help when looking for data is a practice that is especially employed by less experienced users. Differences in goals, requirements, and problems along the experience spectrum could also be confirmed (hypotheses 2a, 2b, 2c). Experienced data users have more ambitious goals (e.g., scientific publication), while less experienced users have rather basic goals (e.g., learning how to work with survey data). Higher experience goes along with advanced requirements (e.g., high quality data, international data), lower experience is associated with less advanced requirements (e.g., data must be easy to understand). Experience is negatively correlated with general problems (e.g., where to find data) and positively correlated with very specific problems (lack of data quality). A very clear relationship could be made out between experience and community involvement (hypothesis 3, the community involvement hypothesis). With regard to the problem solving hypothesis (hypothesis 4), it was found that for specific problems, community involvement supports problem solving. In particular, being denied access to data leads the more involved data users to value contacting principal investigators as a problem solving strategy. This nexus supports the theory's assumption that very experienced researchers use their community involvement to engage in "data dealing" (subchapter "C.3.2.1 Key Findings and

Hypotheses"). On a less spectacular note, they also benefit from direct contacts with principal investigators when they experience problems with insufficient documentation or information on the data. Data users who are less involved in the community may lack this option and thus be prevented from using the data or using it as intended. It seems appropriate here to point out the differences of looking for journal articles and looking for datasets. Both these problems – being denied access and insufficient documentation – are specific to data use and have long been overcome for the case of searching and accessing journal articles.

The relevance of community involvement for data seeking behaviour is conclusive. The results from the quantitative study require several adaptions of the model of information seeking behaviour of survey data users. But overall, the theory of goal-oriented problem solving was successfully exemplified. The consolidated model is presented in the following section.

2.2 A Model of Data Users‘ Information Seeking Behaviour

The general sketch of dependencies between experience, community involvement, and different aspects of information seeking behaviour (goals, requirements, problems, practices) as it was represented in the initial model of problem-solving by community involvement (Figure 36) turned out to be useful and was kept for the consolidated model. The quantitative study led to a more concise, consolidated model of survey data users' information seeking behaviour (Figure 37). The consolidated model retains the understanding that survey data seeking behaviour depends on experience as well as community involvement. Survey data users' experience ranges from low for outsiders to high for experts in the field of survey research. In parallel, their community involvement ranges from low to high. The positive correlation between experience and community involvement could be shown in the quantitative study. Along the spectrum of experience and community involvement, survey data users have different goals, requirements, sources of known data, practices of finding data, problems, and problem solving strategies. In the diagram, these variables are depicted on the left. In the quantitative study, correlations of these variables with experience and/or community involvement were investigated. For each of the investigated variables, the diagram lists those manifestations (or items) that showed significant correlations in the quantitative analysis. The items on the left side of the

spectrum showed negative correlations with experience and/or community involvement whereas the items of the right showed positive correlations.

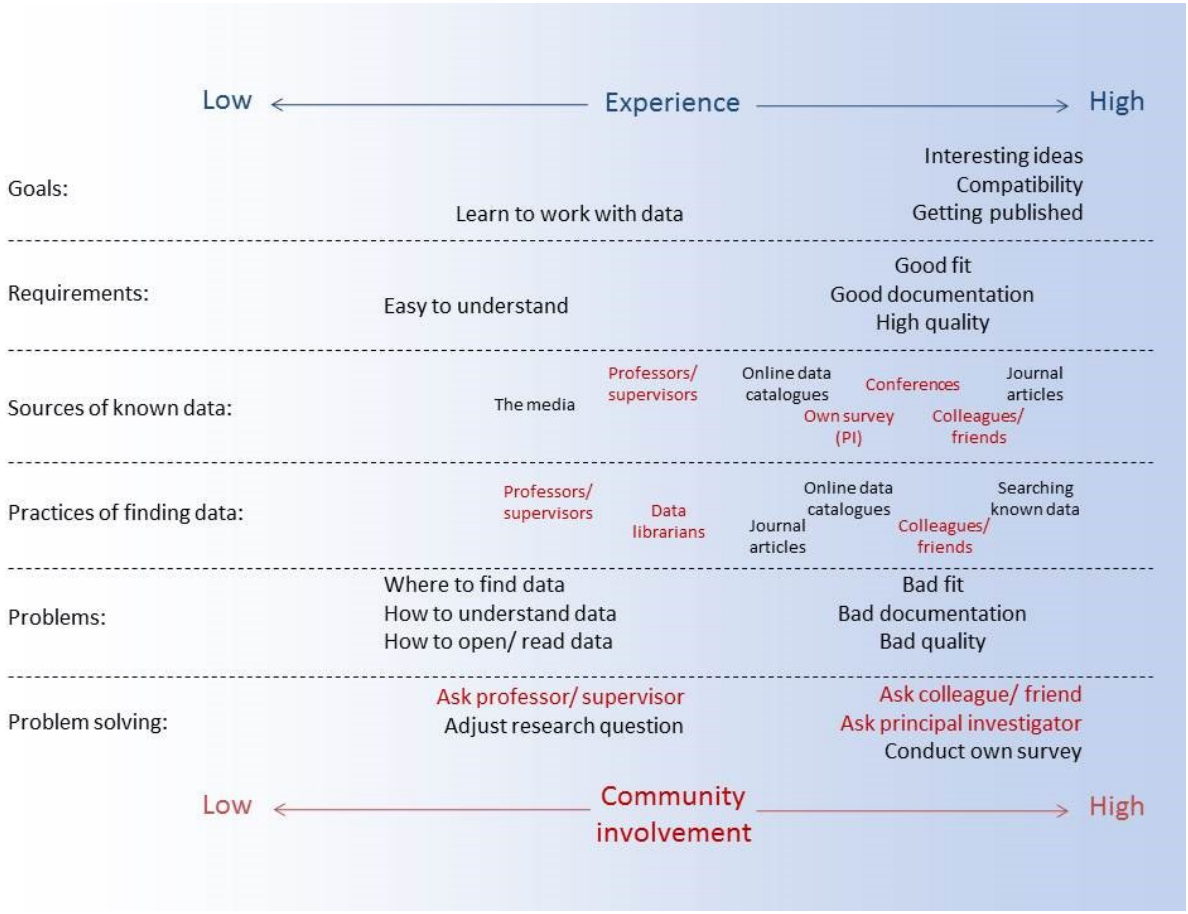


Figure 37 Consolidated model of survey data users' information seeking behaviour

As expected on the grounds of the qualitative study, the **goals** of less experienced users are less ambitious. The only negative correlation with experience was found for the goal of practicing or learning how to work with data. More experienced users have more ambitious goals such as getting published and use of existing measures, e.g. from established survey programmes. Use of existing measures ensures compatibility of results. The positive correlation of experience with the goal of replicating results was weak, compared to the use of data for scientific publication, for teaching and for the use of existing measures. This suggests that the goal of coming up with interesting ideas and reaching interesting results (original research) is important for researchers with much experience. This result is in line with the assumption from the qualitative study that experts aim at innovative and

outstanding findings. With regard to the **requirements** it was found that the requirement of data being easy to understand correlates negatively with experience, which is in line with the findings from the qualitative study. While it could not be confirmed in the quantitative study that the requirement of using data from a specific (popular) survey is associated with more or less experience, it could be shown that with more experience, the requirements of high data quality and good documentation are important. However, there was no significant correlation of experience with the requirement of exclusiveness of data (that have not been analysed before), which had been expected after the qualitative study. For the **sources of known data** as well as **practices of finding data**, the analysis showed that with growing experience and community involvement, the diversity of used sources and practices is increasing. Users on the left side of the spectrum rely on the media, professors/supervisors, and data librarians. The more experienced and involved users on the right side of the spectrum use a whole range of sources and practices, with the use of journal articles being the most important source and practice. Departing from the expectations formulated after the qualitative study, the use of data catalogues to find data is positively correlated with experience and is thus not a typical practice of novice users. Consulting with intermediaries to find data is important for novice users as well as for more experienced users, as it was expected. However, the circle of intermediaries or the *invisible college* of more experienced data users tends to be more diverse. For the more experienced users, the **problems** with regard to finding survey data mirror their requirements, as it is depicted in the diagram (e.g. for the requirement "good fit", there is the problem of "bad fit"). The less experienced survey data users have very basic problems, such as how to open or read a dataset, as it was expected beforehand. Finally, the diagram includes the finding from the quantitative study that with regard to **problem solving**, more community involvement is associated with strategies that involve personal contact, in particular with approaching the principal investigators who have created the data in question.

In comparison, the initial model of problem-solving by community involvement contains several uncertainties, especially with regard to the characteristics and practices of data seeking. Various characteristics could not be placed with certainty with relation to experience beforehand. For example, the practice of web searching was included on the far left of the spectrum (low experience), even though it had been expected that this practice

should be relevant at all levels of experience. The quantitative study underscores this assumption which means that web searching is not a practice that is dependent on experience. Also, personal contacts as a means of information seeking were represented inadequately on the spectrum of experience in the initial model. The initial model depicted intermediaries at an advanced level of experience, even though it was to be expected, that intermediaries were important at every level of experience. Here, the consolidated model clearly benefits from the quantitative study's results that led to a more differentiated picture of the use of personal contacts with regard to experience. The model demonstrates the diversity of personal contacts that are employed in data seeking in relation to experience and community involvement (red entries in Figure 37). Additionally, the quantitative study underscores the importance of the practice of chaining from journal articles when looking for data, in particular for more experienced users. This turned out to be the most important practice of data seeking that had been underestimated in the initial model. Likewise, not all requirements that had been included in the initial model could be confirmed to be dependent on experience. This refers to the requirements of data coming from a specific survey, complex data (e.g. data coming from longitudinal surveys), data that are compatible with other data, and data that have not been analysed a lot before. The reason why no correlations could be found here may be that the answer options were not ideal (see subchapter D. 6.2.2). Ultimately, the requirements that showed correlations with experience form a very plausible equilibrium with the problems that data users experience when they are trying to find suitable data. For example, users with less experience require data that are easy to understand and experience problems with understanding data. Users with more experience require good documentation and experience the opposite.

The consolidated model can serve as a conceptual starting point for further research into survey data seeking behaviour. Furthermore, it lists core concepts in a concise way that should help the further discourse on development of research data infrastructure and services. First ideas and recommendations on possible developments are outlined in the following paragraphs.

3. Recommendations for Research Data Infrastructure Design: Meeting Immediate Challenges

As a whole, research data infrastructure is complex, diverse and distributed. The initiatives on international and national level to improve, connect, and integrate research data services are manifold and are still growing in numbers. To only name a few of these initiatives, the recent years have seen the emergence of the international Research Data Alliance and its regional subsidiaries[24], the establishment of the European Open Science Cloud[25], and the development of the FAIR data principles[26]. Meanwhile, research data production and reuse continue to happen independently and unconnected all over the world and in all disciplines. If we intend to build accessible research data services, a crucial first step is to look at the data from the perspectives of the communities that create these data, work with these data and reuse these data. The present study has taken this perspective for the case of survey data as it is used by the community of survey researchers.

One result of this study is that it has revealed the community's requirements with regard to data seeking. Experienced researchers in particular need high quality data that are well documented. This kind of data is already reliably produced by large survey programmes that are conducted on national and international level. The data are documented and distributed by specialized institutions such as social science data archives. For decades, these archives have provided added value for demanding researchers. They also cooperate on an international level to develop standards that make their data reusable in wider contexts. For example, they have decisively contributed to the development of the DDI standard[27] for survey data documentation. If anything, the present study stresses the importance of these large survey programmes, of the distributing and archiving institutions, and of their cooperation. The expert interviews have provided insight into the complexity of this kind of data and into the tedious work that is necessary to make them available at a reasonable quality. And if less experienced researchers indicate that they require data that are easy to understand, it means that working with this kind of data is indeed challenging. From the expert interviews it became clear that just being able to open or read a dataset does not mean that people are actually able to make good analyses. Knowing how to really read a

[24] https://rd-alliance.org/, accessed October 5, 2020.
[25] https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud, accessed October 5, 2020.
[26] https://www.go-fair.org/fair-principles/, accessed October 5, 2020.
[27] http://www.ddialliance.org/, accessed October 5, 2020.

dataset includes not only understanding the documentation and information material, but also actually using the information that these materials provide (for instance, information on sampling and weighting). The specific documentation is of paramount importance for the reuse of survey data, which is especially challenging for non-disciplinary data services. Hence, an important recommendation that comes out of this study is to keep up the strength of the already existing research data services in the social sciences while striving to achieve more integrated, transdisciplinary services and standards. Integration must not result in levelling down and in reduced quality.

Another major result of this study is that impersonal ways of finding data are not necessarily the most successful. Apparently, survey data users primarily know data from journal articles – not from catalogues of data archives, not from other research data repositories, and not even from searching the web. For the development of research data services this means that survey data are currently not described in a way that makes them easy to find. If we suppose that researchers are searching for data on topics or concepts, we have to annotate the online records of these data with respective metadata – ideally in a standardized way. Currently, researchers have to screen journals to find data or oftentimes, if they already have an idea what survey might fit, have to open and search multiple questionnaires, codebooks, and reports to find out whether a dataset contains data on a specific topic of interest. This is extremely inefficient and will become even more problematic with increasing amounts of data. The situation is further complicated by the fact that while individual data catalogues and repositories hold a broad range of interesting, well documented data, their content has too little visibility on the web. There are too many isolated data silos that don't interoperate and that are using technology that prevents search engines from indexing them. The analysis of the present study has shown that only for more experienced researchers, catalogues and repositories play a role in finding data. Creating meta search engines that integrate or crawl various catalogues seems like a good idea, but in the end, they are just some other information hub that users don't know about. Searching the web for data is an important practice across all levels of experience, which means that research data should be easily findable by web search. In the current situation, too many datasets are still hidden in the Deep Web (He et al. 2007; Chapman et al. 2019), which makes it

unnecessarily hard to retrieve them. Using schema.org vocabularies[28], semantic web standards[29] and tools like semantic sitemaps (Cyganiak et al. 2008) when creating data catalogues and repositories would be a good start to make that happen.

The results also show that the "wheres" and "hows" of finding data need to be even more present in the education of young researchers. This can be drawn from the fact that the second and third most mentioned sources of known data are personal contacts (professors/supervisors and colleagues/friends). This means that the community provides its members with information on available and suitable data. Not the web does this – but the community. This means that, apart from making survey data more findable over the web, the communities' knowledge transfer activities need more support. This refers to offline activities such as survey-specific workshops as well as online activities, in particular on social media. Another interesting finding in this regard is that more experienced survey data users don't turn to data librarians and other data specialists if they are looking for data. These reference persons are mainly consulted by less experienced users. Researchers with more experience might use other services provided by data librarians, but if they need help when looking for data, they turn to their peers. This may be a pointer to data librarians to focus their reference activities on novice data users and maybe support professors and other lecturers in their efforts to educate students in data use.

Finally, the fact that survey researchers are looking for data in relevant journals cannot be dismissed, just because of the plans to successfully improve online data retrieval. Admittedly, this practice of looking for data may only be a popular workaround, because data retrieval on the web is so disappointing. However, finding data through journal articles has its own benefits for several possible reasons. For instance, learning about a dataset in the context of someone else's study can create a better understanding of the data and thus help to make a relevance judgement. Hence, it is necessary to tread both paths: improve findability of survey data on the web and develop services that support finding data in journal articles. The linkage of data and journal articles is already being addressed in several projects[30], but it needs more attention. Metadata records of survey data should include

[28] https://schema.org/, accessed October 5, 2020.
[29] https://www.w3.org/standards/semanticweb/, accessed October 5, 2020.
[30] For instance, the InFoLis project (http://infolis.github.io/, accessed October 5, 2020) and the Scholix service (http://www.scholix.org/, accessed October 5, 2020).

bibliometric information as comprehensively as possible. Authors of empirical research should comply with data citation standards and, by all means, use persistent identifiers when citing datasets. There are promising technological approaches to improve linking between data and journal articles, but further commitment of the community is needed. Not only will citing data improve findability of data; it can also contribute to the long needed system of credit for data sharing. Generating credit for data collection and data sharing will hardly work if datasets are not properly cited. The technology and respective standards are available – now it is up to the communities to install policies that can actually make a difference.

4. Conclusion and Outlook

The primary research contribution of the present study is that it highlights the role of community involvement in data seeking. Survey data communities are an important determinant in survey data users' information seeking behaviour. Community involvement facilitates data seeking and reduces problems or barriers. Community involvement is especially helpful when looking for data – it is less important when looking for literature. If researchers are looking for journal articles or books, they rarely need assistance anymore; web searches will do the trick for almost any request that they might have. To put it the other way around: if they cannot find a paper on a specific topic, they just assume that it doesn't exist. If looking for survey data, they also try to find something on the web; but they may be more successful if they just ask their professors, supervisors or peers. Why is that? There seem to be two major reasons. One is that research data are a complex information source. They are not as easily described and hence retrieved as literature. The second reason is that there is a lack of standardization and completion that contribute to this problem.

Will survey data communities lose their importance if survey data documentation catches up with web standards and if survey data documentation is standardized across institutions and countries? For sure, these measures would go a long way to address problems of insufficient documentation and quality. But problems such as access restriction or compatibility issues will persist. As soon as problems supersede the mere finding of data towards usability judgement, problem solving will always be supported by community involvement.

Apart from these primary results, the present study has yielded further secondary insights. For instance, the qualitative interviews cast light on the whole spectrum of contributions that different people make to provide the research community with reusable data. Apart from primary investigators, there are many other people involved, each with different responsibilities. Associated researchers, research support staff, data managers, and people who work in the field institutes belong to this group that co-create reusable research data. Data curators, data archivists, data librarians and other people who work in data service fulfil a particular important role when it comes to distribute data for reuse and enable secondary data users to reuse them. The interviews in the present study have shown that for the most important datasets, all these people work together and share responsibilities in order to make data sharing a reality. The interviews with data service experts have demonstrated that making data reusable is demanding, sometimes tedious work. In particular, producing documentation for complex datasets is an ambitious task that develops in exchange with different stakeholders.

With regard to data sharing, the quantitative study has produced quite some data that has not been analysed in detail in the present analysis. In general, it has been shown that about 74 percent of the respondents had already collected data and about 53 percent of these had shared these data. These numbers are particularly interesting, because of the high educational and professional level of the respondents in the sample (about 50 percent with a doctoral or higher degree; about 41 percent university or college professors in the sample). With the data collected in this study, further analyses can be made regarding the data sharing behaviour of 534 researchers with experience above average in a data intensive field.

Another interesting secondary result of the present study that has not been analysed further at this point refers to the relevance of social media for data seeking behaviour. At the outset of the study, it was deemed plausible that developments in online communication furthered possibilities of exchange between researchers and intermediaries and thus played an increasingly important role with regard to data seeking practices. However, the results of the quantitative study suggest that social media channels are not among the most important sources when looking for data. Only about 6 percent of the respondents indicated that they had used social media to find reusable data. As a source for known data, social media seems

to be more important – about 13 percent of respondents indicated to know the surveyed datasets from social media channels.

Finally, with regard to the research design, the mixed methods approach proved to be well suited to investigate the research question. The same design could be followed for similar investigations with regard to other data types, in other disciplines, or across multiple disciplines. In particular, the survey instrument (questionnaire) can be reused in other studies with similar research questions. In general, all data that have been collected in the present study have been archived together with the supplementary material and are reusable for further research.

Alongside the presented results, this study has left some questions unaddressed and also yielded new questions. The quantitative study did not factor in all results of the qualitative study and several relevant aspects of data users' information seeking behaviour remain untested. Questions that should be addressed with further research are:

- *Who are members of survey data communities and how do they form?* Dataset communities have emerged early in the conduct of the qualitative study as a core concept in the grounded theory of problem solving by community involvement. The participants in the expert interviews described how dataset communities emerge and persist, because knowledge of them is handed down from senior researchers to junior researchers or shared between peers. Students tend to revisit surveys that they have been introduced to during their education. Young researchers gain knowledge of data infrastructure services such as data repositories or data archives. More advanced researchers find new or other datasets through interaction with peers, for example, at conferences. Some researchers repeatedly work with data from one survey and don’t look for alternatives throughout their careers. For these users and other interested researchers, large survey programmes offer services such as exclusive conferences or “meet the data” workshops for users of their data. Dataset communities that form around a survey or a collection of datasets are made up from people who play any role in the preparation, distribution, finding, and use of these datasets. The community comprises the survey’s principal investigators and other primary researchers, people in the field institutes (interviewers, coordinators

etc.), data managers, data curators, data librarians, and the data users. Within a community, the survey's datasets as well as complementary information on these datasets (documentation) are produced, shared and used. Some community members play different roles at once, for example, they are creators/co-creators and users of a dataset. Sometimes secondary users apply themselves in data improvement, for example, when they detect and report errors in the data. And sometimes, principal investigators make suggestions for improvement to data curators on behalf of secondary users. The same person can have different roles in different survey communities.

- *How do members of survey data communities share their responsibilities and how do they interact to enable research data sharing?* An interesting question to explore in further research would be how community members congregate, interact, communicate, and share responsibilities. Of particular interest here are mechanisms or incentives that further or impede involvement in survey data communities. From this information, support mechanisms for community structures could be developed. A particularly suitable method to investigate dataset communities seems to be network analysis. It would be interesting, for instance, to conduct a scholarly network analysis (White 2011) of the community of a well-known and frequently used survey programme such as the German Allbus. Going further, it would be interesting to analyse in depth how community members interact when they are sharing, creating, improving, distributing and reusing datasets. This analysis could, for instance, be based on an adapted model of collaborative information seeking (CIS) as presented by Chirag Shah (Shah 2014).
- *Are survey researchers with strong community involvement more successful?* To investigate this question, a mixed methods approach is appropriate. The respondents could be sampled from authors of survey research in relevant journals of disciplines with a focus on quantitative empirical research (such as sociology, political science, economics, education, and psychology), a sampling method that is frequently used in the field (e.g. Yoon 2017; Zenk-Möltgen et al. 2018). First, a quantitative study should survey researchers' community involvement. Measurements of community involvement and experience can be adopted from the present study. In addition to the survey, a bibliometric analysis would be performed to estimate the respondents'

success with publications. A combination of the results of both the survey and the bibliometric study could be used to find answers to the question whether strong community involvement influences success in scholarly work.

- *How are survey researchers searching the web for data?* While the present study investigated data seeking behaviour from a more contextual point of view, it merely touched on specific questions of data searching or data retrieval. It has been pointed out that finding survey data by searching the web is difficult and complicated by lack of standardized indexing. Improvement in this area could be informed by knowledge about users' actual search behaviours. For instance, it would be helpful to have empirical knowledge about terminology and semantics used by researchers who are searching the web for data. There has been quite some research in that area already (Groth et al. 2018). With regard to survey data, it would be particularly interesting to investigate how the well documented records from long established data archives are retrievable by searching the web instead of searching the distributed catalogues and portals that not everybody in every discipline might know. It would also be interesting to investigate, whether retrieval practices differ when researchers search catalogues and repositories as opposed to general web search engines.
- *Why and how are data users searching journal articles to find suitable data?* It remains one of the most interesting findings of this study that searching journal articles for suitable data is a very important, if not most important strategy of data seeking for survey researchers. The reasons behind the broad use of this strategy as well as the practices involved in scanning journals for data are worth a closer look. Qualitative methods such as interviews, focus groups, and user studies with thinking aloud techniques would be appropriate to investigate these questions. The results of these studies could also be related to the concepts of community involvement and experience that have been developed in the present study.
- *Are data communities relevant for data seeking behaviour in other disciplines?* Recent research carried out by Kathleen Gregory and colleagues (Gregory, Cousijn, et al. 2019; Gregory, Groth, et al. 2019) would suggest so. In a study combining a bibliometric analysis of current research on data search and qualitative interviews with 22 users of the Elsevier data search portal, coming from multiple research fields, Gregory, Cousijn et al. studied various behaviours surrounding data seeking. The

authors interpret data seeking as a contextual, socio-technical practice and emphasise the role of social interactions when locating data for reuse, for example with colleagues, supervisors, principal investigators, and support staff. For some researchers, the authors found, “personal connections are the most efficient and accurate route to data search” (Gregory, Cousijn, et al. 2019, 10). Similar to what was found in the qualitative part of the present study, Gregory, Cousijn et al. also found that there are exclusive communities that enable their members to access restricted data. In another recent investigation, Gregory, Groth et al. surveyed about 1.700 data users from different disciplines, recruited from indexed authors from the Elsevier Scopus database (Gregory, Groth, et al. 2019). Regarding data discovery practices, the authors found that respondents named literature as the most important source for finding data, which is perfectly in line with the results of the present study. Another joint result is that apart from using literature to find data, researchers turn to their personal contacts ("communities of data seekers") to find and access data (Gregory, Groth, et al. 2019, 23). The authors conclude that data seeking and accessing are mediated processes - mediated either by literature or social connections. To follow up these results, the model of problem solving by community involvement as it has been developed in the present study could be tested with regard to other disciplines. The results from Gregory and colleagues suggest that this could be a promising approach.

- *Finding, sharing and talking about data on social media – does that happen?* In the quantitative part of this study, survey data users were asked whether they used social media to find data, to share their data, and to get help with data. In general, the numbers on all these questions were quite low. The present study did not investigate these data any further, for example with regard to with regard to background variables. Further analyses in that direction could lead to interesting findings. For example, is the use of social media more important for a certain age group? Research on academic activities on social media would suggest so (Mohammadi et al. 2018). In her 2017 qualitative study of data users from the fields of public health and social work, Ayoung Yoon found that in the whole process of reusing data, from finding to analyzing data, researchers communicated with various people to receive support (Yoon 2017). Yoon describes how the participants

explained that they regularly had a "data talk" with other researchers (Yoon 2017, 465). The present study asked respondents if they had used social media to solve problems of finding or accessing data, which could be interpreted as a specific case of "data talk". The data collected here could be analyzed further in that direction.

The long line of possible research questions suggests that it we still have a long road to go until "shared data [can] be as easily decoded as a shared library book" (David 1991, 93).

References

Agarwal, Naresh Kumar. 2017. *Exploring Context in Information Behavior: Seeker, Situation, Surroundings, and Shared Identities*. Morgan & Claypool Publishers.

Allen, Thomas. 1969. 'Information Needs and Uses'. *Annual Review of Information Science and Technology* 4: 3–29.

Alliance of German Science Organisations. 2010. 'Principles for the Handling of Research Data'. https://gfzpublic.gfz-potsdam.de/rest/items/item_4507890_1/component/file_4507888/content (October 5, 2020).

American Association of Public Opinion Research. 2016. 'Standard Definitions. Final Dispositions of Case Codes and Outcome Rates for Surveys. Revised 2016'. https://www.aapor.org/AAPOR_Main/media/publications/Standard-Definitions20169theditionfinal.pdf (October 5, 2020).

Athukorala, Kumaripaba et al. 2014. 'Information-Seeking Behaviors of Computer Scientists: Challenges for Electronic Literature Search Tools'. *Proceedings of the American Society for Information Science and Technology* 50(1): 1–11.

ATLAS.ti Scientific Software Development. 2012. *ATLAS.Ti. Version 7*. Berlin: ATLAS.ti Scientific Software Development GmbH.

Azama, Mohammad, and Rahmattolah Fattahi. 2011. 'Matching the Databases' User Interface with Ellis' Model of Information Seeking Behavior: A Qualitative Study'. In *New Trends in Qualitative and Quantitative Methods in Libraries. Selected Papers Presented at the 2nd Qualitative and Quantitative Methods in Libraries*, eds. Anthi Katsirikou and Christos Skiadas. Singapore: World Scientific, 287–96.

Bates, Marcia J. 2004. 'Information Science at the University of California at Berkeley in the 1960s: A Memoir of Student Days'. *Library Trends* 52(4): 683–701.

———. 2010. 'Information Behavior'. In *Encyclopedia of Library and Information Sciences*, eds. Marcia J. Bates and Mary Niles Maack. New York: CRC Press, 2381–91.

Bawden, David, and Lyn Robinson. 2013. 'No Such Thing as Society? On the Individuality of Information Behavior'. *Journal of the American Society for Information Science and Technology* 64(12): 2587–90.

Bernard, H. Russel. 2013. *Social Research Methods. Qualitative and Quantitative Approaches*. 2nd edition. Thousand Oaks: SAGE.

Blandford, Ann, and Simon Attfield. 2010. 'Interacting with Information'. *Synthesis Lectures on Human-Centered Informatics* 3(1): 1–99.

Blue Ribbon Task Force on Sustainable Digital Preservation and Access. 2010. 'Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information. Final Report of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access'. https://www.cs.rpi.edu/~bermaf/BRTF_Final_Report.pdf (October 5, 2020).

Borgman, Christine L. 2007. *Scholarship in the Digital Age. Information, Infrastructure, and the Internet*. Cambridge: MIT Press.

———. 2010. *Research Data: Who Will Share What with Whom, When, and Why?* . RatSWD Working Paper. http://www.ratswd.de/download/RatSWD_WP_2010/RatSWD_WP_161.pdf (October 5, 2020).

———. 2012. 'The Conondrum of Sharing Research Data'. *Journal of the American Society for Information Science and Technology* 63(6): 1059–78.

Bouazza, Abdelmajid. 1989. 'Information User Studies' ed. Allen Kent. *Encyclopedia of Library and Information Science. Vol. 44, supp. 9*: 144–64.

Bronstein, Jenny. 2007. 'The Role of the Research Phase in Information Seeking Behavior of Jewish Studies Scholars: A Modification of Ellis' Behavioural Characteristics'. *Information Research* 12(3). http://informationr.net/ir/12-3/paper318.html (October 5, 2020).

Brown, Mary E. 1991. 'A General Model of Information-Seeking Behavior'. In , 9–14.

Bryant, Antony, and Kathy Charmaz. 2007. 'Grounded Theory in Historical Perspective: An Epistemological Account'. In *The SAGE Handbook of Grounded Theory*, eds. Antony Bryant and Kathy Charmaz. London: SAGE, 31–57.

Bryman, Alan. 2012. *Social Research Methods*. 4th edition. New York: Oxford University Press.

Bulmer, Martin, Patrick J. Sturgis, and Nick Allum. 2009. 'Editors' Introduction'. In *Secondary Analysis of Survey Data*, eds. Martin Bulmer, Patrick J. Sturgis, and Nick Allum. Los Angeles: SAGE, XXIII–XXVI.

Caidi, Nadia, Danielle Allard, and Lisa Quirke. 2010. 'Information Practices of Immigrants'. *Annual Review of Information Science and Technology* 44: 493–531.

Callegaro, Mario, Katja Lozar Manfreda, and Vasja Vehovar. 2015. *Web Survey Methodolgy*. London: Sage.

Carlson, Samuelle, and Ben Anderson. 2007. 'What Are Data? The Many Kinds of Data and Their Implications for Data Re-Use'. *Journal of Computer-Mediated Communication* 12(2): 635–51.

Case, Donald O. 2012. *Looking for Information: A Survey of Research on Information Seeking, Needs and Behavior*. Bingley: Emerald Group Publishing.

Case, Donald O., and Lisa M. Given. 2016. *Looking for Information: A Survey of Research on Information Seeking, Needs, and Behavior*. Emerald Group Publishing Limited.

Chapman, Adriane et al. 2019. 'Dataset Search: A Survey'. *The VLDB Journal*.

Charmaz, Kathy. 2005. 'Grounded Theory in the 21st Century. Applications for Advancing Social Justice Studies'. In *The Sage Handbook of Qualitative Research*, eds. Norman K. Denzin and Yvonna S. Lincoln. Thousand Oaks: SAGE, 507–37.

———. 2008. 'Reconstructing Grounded Theory'. In *The SAGE Handbook of Social Research Methods*, eds. Pertti Alasuutari, Leonard Bickman, and Julia Brannen. London: SAGE.

———. 2014. *Constructing Grounded Theory*. 2nd edition. London: Sage.

Choo, Chun Wei, Brian Detlor, and Dan Turnbull. 2000. 'Information Seeking on the Web: An Integrated Model of Browsing and Searching'. *First Monday* 5(2). http://journals.uic.edu/ojs/index.php/fm/article/view/729 (October 5, 2020).

Circle Systems. 2015. *Stat/Transfer. Version 14*. Seattle: Circle Systems Inc.

Clark, Rich, and Marc Maynard. 1998. 'Research Methodology: Using Online Technology for Secondary Analysis of Survey Research Data - "Act Globally, Think Locally"'. *Social Science Computer Review* 16(1): 58–71.

Clubb, Jerome M., Erik W. Austin, Carolyn L. Geda, and Michael W. Traugott. 1985. 'Sharing Research Data in the Social Sciences'. In *Sharing Research Data*, eds. Stephen E. Fienberg, Margaret E. Martin, and Miron L. Straf. Washington, D.C.: National Academy Press, 39–88.

Corti, Louise. 2004. 'Data Archives' eds. Michael S. Lewis-Beck, Alan Brymann, and Tim Futing Liao. *The Sage Encyclopedia of Social Science Research Methods* 1: 234–36.

Courtright, Christina. 2007. 'Context in Information Behavior Research'. *Annual Review of Information Science and Technology* 41: 273–306.

Creswell, John W. 2013. *Qualitative Inquiry and Research Design. Choosing among Five Approaches*. 3rd edition. Thousand Oaks: Sage.

———. 2014. *Research Design. Qualitative, Quantitative, and Mixed Methods Approaches*. 4th edition. Thousand Oaks: Sage.

Creswell, John W., and Vicky L. Plano Clark. 2011. *Designing and Conducting Mixed Methods Research*. 2nd edition. Thousand Oaks: Sage.

Cronin, Blaise. 1982. 'Invisible Colleges and Information Transfer. A Review and Commentary with Particular Reference to the Social Sciences'. *Journal of Documentation* 38(3): 212–36.

Curty, Renata G. 2015. 'Beyond "Data Thrifting": An Investigation of Factors Influencing Research Data Reuse in the Social Sciences. Dissertations - ALL; Paper 266.' http://surface.syr.edu/etd/266/ (October 5, 2020).

———. 2016. 'Factors Influencing Research Data Reuse in the Social Sciences: An Exploratory Study'. *International Journal of Digital Curation* 11(1): 96–117.

Curty, Renata G., and Jian Qin. 2014. 'Towards a Model for Research Data Reuse Behavior'. *Proceedings of the American Society for Information Science and Technology* 51(1): 1–4.

Curty, Renata G., Ayoung Yoon, Wei Jeng, and Jian Qin. 2016. 'Untangling Data Sharing and Reuse in Social Sciences'. *Proceedings of the American Society for Information Science and Technology* 53(1): 1–5.

Cyganiak, Richard et al. 2008. 'Semantic Sitemaps: Efficient and Flexible Access to Datasets on the Semantic Web'. In *The Semantic Web: Research and Applications. ESWC 2008*, Lecture Notes in Computer Science, eds. Sean Bechhofer, Manfred Hauswirth, Jörg Hoffmann, and Manolis Koubarakis. Berlin, Heidelberg: Springer, 690–704. http://link.springer.com/10.1007/978-3-540-68234-9_50 (October 5, 2020).

Dahinden, Urs. 2013. 'Methoden Empirischer Sozialforschung Für Die Informationspraxis'. In *Grundlagen Der Praktischen Information Und Dokumentation. Handbuch Zur Einführung in Die Informationswissenschaft Und -Praxis*, eds. Rainer Kuhlen, Wolfgang Semar, and Dietmar Strauch. Berlin, Boston: De Gruyter, 126–35.

David, Martin. 1991. 'The Science of Data Sharing. Documentation'. In *Sharing Social Science Data. Advantages and Challenges*, Sage focus editions, ed. Joan E. Sieber. Newbury Park: Sage, 91–115.

Dervin, Brenda. 2003. 'Given a Context by Another Name: Methodological Tools for Taming the Unruly Beast'. In *Sense-Making Methodology Reader*, eds. B. Dervin, L. Foreman-Wernet, and E. Lauterbach. Cresskill: Hampton Press, 111–32.

Deutsche Forschungsgemeinschaft. 2015. 'DFG Guidelines on the Handling of Research Data'. https://www.dfg.de/download/pdf/foerderung/antragstellung/forschungsdaten/guidelines_research_data.pdf (October 5, 2020).

Dillman, Don A., Jolene D. Smyth, and Leah Melani Christian. 2009. *Internet, Mail, and Mixed-Mode Surveys. The Tailored Design Method*. 3rd edition. Hoboken: Wiley.

Economic and Social Research Council. 2018. 'ESRC Research Data Policy. November 2014. Last Updated 2018.' https://esrc.ukri.org/files/about-us/policies-and-standards/esrc-research-data-policy/ (October 5, 2020).

Ellis, David. 1989. 'A Behavioural Approach to Information Retrieval System Design'. *The Journal of Documentation* 45(3): 171–212.

———. 2011. 'The Emergence of Conceptual Modelling in Information Behaviour Research'. In *New Directions in Information Behaviour*, Library and Information Science, eds. Amanda Spink and Jannica Heinström. Bingley: Emerald, 17–35.

Ellis, David, Deborah Cox, and Katherine Hall. 1993. 'Comparison of the Information Seeking Patterns of Researchers in the Physical and Social Sciences'. *Journal of Documentation* 49(4): 356–69.

Faniel, Ixchel M., Julianna Barrera-Gomez, Adam Kriesberg, and Elizabeth Yakel. 2013. 'A Comparative Study of Data Reuse Among Quantitative Social Scientists and Archaeologists'. In *IConference 2013 Proceedings*, , 797–800.

Faniel, Ixchel M., Adam Kriesberg, and Elizabeth Yakel. 2012. 'Data Reuse and Sensemaking among Novice Social Scientists'. *Proceedings of the Association for Information Science and Technology (ASIST)* 49(1): 1–10.

———. 2016. 'Social Scientists' Satisfaction With Data Reuse'. *Journal of the Association for Information Science and Technology* 67(6): 1404–16.

Fidel, Raya. 2008. 'Are We There yet? Mixed Methods Research in Library and Information Science'. *Library & Information Science Research* 30: 265–72.

Fisher, Karen E., Sanda Erdelez, and Lynne McKechnie. 2005. *Theories of Information Behavior*. Medford: Information Today.

Fisher, Karen, and Heidi Julien. 2009. 'Information Behavior'. *Annual Review of Information Science and Technology* 43(1): 1–73.

Fitzgerald, Sarah Rose. 2018. ‘Serving a Fragmented Field: Information Seeking in Higher Education’. *The Journal of Academic Librarianship* 44(3): 337–42.

Flechter, John. 1982. ‘Economics’. In *The Social Sciences: The Supply of and Demand for Documentation and Data*, ed. Michael Brittain. London: Rossendale, 16–23.

Folster, Mary B. 1995. ‘Information Seeking Patterns: Social Sciences’. *The Reference Libarian* 23(49–50): 83–93.

Ford, Geoffrey, ed. 1977. *User Studies. An Introductory Guide and Select Bibliography*. Sheffield.

Friedrich, Tanja. 2020a. ‘Looking for Data (Expert Interviews): Information Seeking Behaviour of Survey Data Users. [Transcripts]’. https://doi.org/10.7802/1.1943 (October 5, 2020).

———. 2020b. ‘Looking for Data (Online Survey): Information Seeking Behaviour of Survey Data Users. [Dataset]’. https://doi.org/10.7802/1.1953 (October 5, 2020).

Friedrich, Tanja, and Andreas Oskar Kempf. 2014. ‘Making Research Data Findable in Digital Libraries: A Layered Model for User-Oriented Indexing of Survey Data’. In *IEEE/ACM Joint Conference on Digital Libraries*, , 53–56.

Friedrich, Tanja, and Pascal Siegers. 2016. ‘The Ofness and Aboutness of Survey Data: Improved Indexing of Social Science Questionnaires’. In *Analysis of Large and Complex Data*, Studies in Classification, Data Analysis, and Knowledge Organization, eds. Adalbert F.X. Wilhelm and Hans A. Kestler. Cham: Springer, 629–638.

Fry, Jenny. 2006. ‘Scholarly Research and Information Practices: A Domain Analytic Approach’. *Information Processing and Management* 42: 299–316.

Fry, Jenny, and Sanna Talja. 2004. ‘The Cultural Shaping of Scholarly Communication: Explaining E-Journal Use Within and Across Academic Fields’. In *Proceedings of the 67th ASIS&T Annual Meeting*, , 20–30.

Gannon-Leary, Pat, Moira Bent, and Jo Webb. 2007. ‘Researchers and Their Information Needs: A Literature Review’. *New Review of Academic Librarianship* 13(1–2): 51–69.

Ge, Xuemei. 2010. ‘Information-Seeking Behavior in the Digital Age: A Multidisciplinary Study of Academic Researchers’. *College & Reserach Libraries* 71(5): 435–55.

GESIS - Data Archive for the Social Sciences. 2019. ‘Download Statistics GESIS Data Archive. [Dataset]’. http://dx.doi.org/10.4232/1.13222 (October 5, 2020).

Glaser, Barney G., and Anselm L. Strauss. 1967. *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Chicago: Aldine.

Gläser, Jochen, and Grit Laudel. 2010. *Experteninterviews Und Qualitative Inhaltsanalyse Als Instrumente Rekonstruierter Untersuchungen*. 4. Auflage. Wiesbaden: VS Verlag.

Gold, Anna. 2007. ‘Cyberinfrastructure, Data, and Libraries, Part 1’. *D-Lib Magazine* 23(1/2). http://www.dlib.org/dlib/september07/gold/09gold-pt1.html (October 5, 2020).

González-Teruel, Aurora, and M. Francisca Abad-García. 2012. ‘Grounded Theory for Generating Theory in the Study of Behavior’. *Library & Information Science Research* 34(1): 31–36.

Gould, Constance C., and Mark Handler. 1989. *Information Needs in the Social Sciences: An Assessment. Prepared for the Program for Research Information Management of The Research Libraries Group, Inc*.

van der Graaf, M., L. Waaijers, and J. [contributor Davidson. 2011. 'A Surfboard for Riding the Wave: Towards a Four Country Action Programme on Research Data. A Knowledge Exchange Report'. http://www.knowledge-exchange.info/event/riding-the-wave (October 5, 2020).

Gregory, Kathleen, Helena Cousijn, et al. 2019. 'Understanding Data Search as a Socio-Technical Practice'. *Journal of Information Science*: 1–17.

Gregory, Kathleen, Paul Groth, Andrea Scharnhorst, and Sally Wyatt. 2019. 'Lost or Found? Discovering Data Needed for Research'. *arXiv:1909.00464 [cs]*. http://arxiv.org/abs/1909.00464 (October 5, 2020).

Groth, Paul et al. 2018. 'DATA: SEARCH'18 - Searching Data on the Web'. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval - SIGIR '18*, Ann Arbor, MI, USA: ACM Press, 1419–22. http://dl.acm.org/citation.cfm?doid=3209978.3210195 (October 5, 2020).

Groves, Robert M. et al. 2009. *Survey Methodology*. 2nd edition. Hoboken: Wiley.

Guba, Egon G. 1990. *The Paradigm Dialog*. Newbury Park: Sage.

Gutmann, M., K. Schürer, D. Donakowski, and Hilary Beedham. 2004. 'The Selection, Appraisal, and Retention of Digital Social Science Data'. *Data Science Journal* 3: 209–21.

Hargittai, Eszter, and Amanda Hinnant. 2006. 'Toward a Social Framework for Information Seeking'. In *New Directions in Information Behavior*, Information Science and Knowledge Management, eds. Amanda Spink and Charles Cole. Dordrecht: Springer, 55–70.

Harris, Colin. 1982. 'Social Policy and Administration'. In *The Social Sciences: The Supply of and Demand for Documentation and Data*, ed. Michael Brittain. London: Rossendale, 36–45.

He, Bin, Mitesh Patel, Zhen Zhang, and Kevin Chen-Chuan Chang. 2007. 'Accessing the Deep Web - A Survey'. *Communications of the ACM* 50(5): 94–101.

Heim, Kathleen M. 1980. *Social Science Data Archives. A User Study*. Wisconsin-Madison.

Hemminger, Bradley M., Dihui Lu, K.T.L. Vaughan, and Stephanie J. Adams. 2007. 'Information Seeking Behavior of Academic Scientists'. *Journal of the American Society for Information Science and Technology* 58(14): 2205–25.

Herb, Ulrich. 2015. *Open Science in Der Soziologie. Eine Interdisziplinäre Bestandsaufnahme Zur Offenen Wissenschaft Und Eine Untersuchung Ihrer Verbreitung in Der Soziologie*. Glückstadt: Hülsbusch.

Herman, Eti. 2004a. 'Research in Progress: Some Preliminary and Key Insights into the Information Needs of the Contemporary Academic Researcher. Part 1'. *ASLIB Proceedings* 56(1): 34–47.

———. 2004b. 'Research in Progress: Some Preliminary and Key Insights into the Information Needs of the Contemporary Academic Researcher. Part 2'. *ASLIB Proceedings* 56(2): 118–31.

Herring, James E. 2013. 'Constructivist Grounded Theory: A 21st Century Research Methodology'. In *Research Methods: Information, Systems and Contexts*, eds. Kirsty Williamson and Graeme Johanson. Chandos Publishing, 203–18.

Hert, Carol A., and Gary Marchionini. 1997. *Seeking Statistical Information in Federal Websites: Users, Tasks, Strategies, and Design. Final Report to the Bureau of Labour Statistics*. http://ils.unc.edu/~march/blsreport/blsmain.htm (October 5, 2020).

———. 1998. 'Information Seeking Behavior on Statistical Websites: Theoretical and Design Implications'. In *Proceedings of the 61st ASIS Annual Meeting*, , 303–14.

Hey, Tony, Stewart Tansley, and Kristin Tolle, eds. 2009. *The Fourth Paradigm. Data-Intensive Scientific Discovery*. Redmond: Microsoft Research.

Hey, Tony, and Anne Trefethen. 2003. 'The Data Deluge: An e-Science Perspective'. In *Grid Computing: Making the Global Infrastructure a Reality*, eds. Fran Berman, Geoffrey Fox, and Tony Hey. Chichester: Wiley, 809–24.

High level Expert Group on Scientific Data. 2010. 'Riding the Wave. How Europe Can Gain from the Rising Tide of Scientific Data. Final Report of the High Level Expert Group on Scientific Data. A Submission to the European Commission'. https://www.fosteropenscience.eu/sites/default/files/original/831.pdf (October 5, 2020).

Hjørland, Birger. 2000. 'Documents, Memory Institutions and Information Science'. *Journal of Documentation* 56(1): 27–41.

———. 2002. 'Epistemology and the Socio-Cognitive Perspective in Information Science'. *Journal of the American Society for Information Science and Technology* 53(4): 257–70.

———. 2005. 'The Socio-Cognitive Theory of Users Situated in Specific Contexts and Domains'. In *Theories of Information Behavior*, ASIST Monograph Series, eds. Karen Fisher, Sanda Erdelez, and Lynne (E.F.) McKechnie. Medford: Information Today, 339–43.

Hjørland, Birger, and Hanne Albrechtsen. 1995. 'Toward a New Horizon in Information Science: Domain-Analysis'. *Journal of the American Society for Information Science* 46(6): 400–425.

Hunsucker, R. Laval. 2007. 'More Appropriate Information Systems and Sservices Fo the Social Scientist: Time to Put Our Findings to Work'. *Evidence Based Library and Information Practice* 2(4): 95–103.

Huschka, Denis, Claudia Oellers, Notburga Ott, and Gert G. Wagner. 2011. 'Datenmanagement Und Data Sharing: Erfahrungen in Den Sozial- Und Wirtschaftswissenschaften'. In *Handbuch Forschungsdatenmanagement*, eds. Stephan Büttner, Hans-Christoph Hobohm, and Lars Müller. Bad Honnef: Bock + Herchen, 35–48.

Jacob, Rüdiger, Andreas Heinz, and Jean Philippe Décieux. 2011. *Umfrage: Einführung in Die Methoden Der Umfrageforschung*. 2. Auflage. München: Oldenbourg.

Jacoby, JoAnn. 2010. 'Share and Share Alike? Data-Sharing Practices in Different Disciplinary Domains'. In *Social Science Libraries: Interdisciplinary Collections, Services, Networks*, IFLA Publications, eds. Steven W. Witt and Lynne M. Rudasill. Berlin: De Gruyter, 79–94.

Janes, Mark. 2009. *Time to Take Another Bath? A Preliminary Report on the Feasibility of Repeating the INFROSS Study*. http://www.ifla.org/node/1011 (October 5, 2020).

Kern, Dagmar, and Brigitte Mathiak. 2015. 'Are There Any Differences in Data Set Retrieval Compared to Well-Known Literature Retrieval?' In *Research and Advanced Technology for Digital Libraries*, Lecture Notes in Computer Science; 9316, eds. Sarantos Kapidakis, Cezary Mazurek, and Marcin Werla. Cham: Springer, 197–208.

Kiecolt, Jill K., and Laura E. Nathan. 1985. *Secondary Analysis of Survey Data*. Newbury Park: Sage.

Kim, Youngseek, and Jeffrey M. Stanton. 2014. 'Institutional and Individual Influences on Scientists' Data Sharing Behaviors: A Multilevel Analysis'. *Proceedings of the American Society for Information Science and Technology* 50(1): 1–14.

Kommission Zukunft der Informationsinfrastruktur. 2011. 'Gesamtkonzept Für Die Informationsinfrastruktur Für Deutschland'. https://www.hof.uni-halle.de/web/dateien/KII_Gesamtkonzept_2011.pdf (October 5, 2020).

Krampen, Günter, Clemens Fell, and Gabriel Schui. 2011. 'Psychologists' Research Activities and Professional Information-Seeking Behaviour: Empirical Analysis with Reference to the Theory of the Intellectual and Social Organization of the Sciences'. *Journal of Information Science* 37(4): 439–50.

Krosnick, J., and S. Presser. 2010. 'Question and Questionnaire Design'. In *Handbook of Survey Research*, eds. P. Marsden and J. Wright. Bingley: Emerald.

Latcheva, Rossalina, and Eldad Davidov. 2014. 'Skalen Und Indizes'. In *Handbuch Methoden Der Empirischen Sozialforschung*, eds. Nina Baur and Jörg Blasius. Wiesbaden: Springer, 745–56.

Law, Margaret. 2005. 'Reduce, Reuse, Recycle: Issues in the Secondary Use of Research Data'. *IASSIST Quarterly* 29(1): 5–10.

Leckie, Gloria J. 2005. 'General Model of the Information Seeking of Professionals'. In *Theories of Information Behavior*, ASIST Monograph Series, eds. Karen Fisher, Sanda Erdelez, and Lynne (E.F.) McKechnie. Medford: Information Today, 158–63.

Lewis-Beck, Michael S. 2004a. 'Data' eds. Michael S. Lewis-Beck, Alan Brymann, and Tim Futing Liao. *The Sage Encyclopedia of Social Science Research Methods* 1: 234.

———. 2004b. 'Secondary Data' eds. Michael S. Lewis-Beck, Alan Brymann, and Tim Futing Liao. *The Sage Encyclopedia of Social Science Research Methods* 3: 1009.

Lincoln, Yvonna S., Susan A. Lynham, and Egon G. Guba. 2011. 'Paradigmatic Controversies, Contradictions, and Emerging Confluences, Revisited'. In *The Sage Handbook of Qualitative Research*, eds. Norman K. Denzin and Yvonna S. Lincoln. Thousand Oaks: Sage, 97–128.

Line, Maurice B. 1971. 'The Information Uses and Needs of Social Sciences: An Overview of INFROSS'. *ASLIB Proceedings* 23(8).

Lloyd, Michael, and Annemaree Olsson. 2017. 'Being in Place: Embodied Information Practices'. *Information Research* 22(1). http://informationr.net/ir/22-1/colis/colis1601 (October 5, 2020).

Lozar Manfreda, Katja et al. 2008. 'Web Surveys versus Other Survey Modes: A Meta-Analysis Comparing Response Rates'. *International journal of market research* 50(1): 79–104.

Lynch, Clifford A. 2009. 'Jim Gray's Fourth Paradigm and the Construction of the Scientific Record'. In *The Fourth Paradigm. Data-Intensive Scientific Discovery*, eds. Tony Hey, Stewart Tansley, and Kristin Tolle. Redmond: Microsoft Research, 177–83.

Ma, Lai. 2012. 'Some Philosophical Considerations in Using Mixed Methods in Library and Information Science Research'. *Journal of the American Society for Information Science and Technology* 63(9): 1859–67.

Marchionini, Gary. 1995. *Information Seeking in Electronic Environments*. Cambridge: Cambridge University Press.

Marcum, Deanna B., and Gerald George, eds. 2010. *The Data Deluge: Can Libraries Cope with E-Science?* Santa Barbara: ABC-CLIO.

McKenzie, Pamela J. 2003. 'A Model of Information Practices in Accounts of Everyday-Life Information Seeking'. *Journal of Documentation* 59(1): 19–40.

———. 2006. 'Mapping Textually Mediated Information Practice in Clinical Midwifery Care'. In *New Directions in Human Information Behaviour*, Information Science and Knowledge Management, eds. Amanda Spink and Charles Cole. Dordrecht: Springer, 73–92.

Medjedović, Irena. 2014. *Qualitative Sekundäranalyse. Zum Potenzial Einer Neuen Forschungsstrategie in Der Empirischen Sozialforschung*. Wiesbaden: Springer.

Meho, Lokman I., and Helen R. Tibbo. 2003. 'Modeling the Information-Seeking Behavior of Social Scientists: Ellis' Study Revisited'. *Journal of the American Society for Information Science and Technology* 54(6): 570–87.

Mishra, Jyoti, David Allen, and Alan Pearman. 2015. 'Information Seeking, Use, and Decision Making'. *Journal of the Association for Information Science and Technology* 66(4): 662–73.

Mohammadi, Ehsan, Mike Thelwall, Mary Kwasny, and Kristi L. Holmes. 2018. 'Academic Information on Twitter: A User Survey'. *PLOS ONE* 13(5): e0197265.

Nathan, Laura. 2004. 'Secondary Analysis of Survey Data'. In *The Sage Encyclopedia of Social Science Research Methods*, eds. Michael S. Lewis-Beck, Alan Brymann, and Tim Futing Liao. Thousand Oaks: Sage, 1008–9.

National Science Foundation. 2012. 'Proposal and Award Policies and Procedures Guide. Part II - Award & Administration Guide'. http://www.nsf.gov/pubs/policydocs/pappguide/nsf13001/aagprint.pdf (October 5, 2020).

Nielsen, Hans Jørn, and Birger Hjørland. 2014. 'Curating Ressearch Data: The Potential Roles of Libraries and Information Professionals'. *Journal of Documentation* 70(2): 221–40.

Niu, Jinfang. 2009. *Perceived Documentation Quality of Social Science Data*.

Niu, Jinfang, and Margaret Hedstrom. 2008. 'Documentation Evaluation Model for Social Science Data'. *Proceedings of the American Society for Information Science and Technology* 45(1): 11.

———. 2009. ‘Documentation Evaluation Model for Social Science Data: An Empirical Test’. In *Digital Curation: Practice, Promise and Prospects. Proceedings of DigCCurr2009*, eds. Helen R. Tibbo, C. Hank, C. A. Lee, and R. Clemens. Chapel Hill, 125–29.

Niu, Xi et al. 2010. ‘National Study of Information Seeking Behavior of Academic Researchers in the United States’. *Journal of the American Society for Information Science and Technology* 61(5): 869–90.

Paisley, William J. 1965. *The Flow of (Behavioral) Science Information: A Review of the Research Literature*. Stanford, Calif.: Stanford University, Institute for Communication Research.

Palmer, Carole L., and Melissa H. Cragin. 2008. ‘Scholarship and Disciplinary Practices’. *Annual Review of Information Science and Technology* 42(1): 163–212.

Palmer, Carole L, Lauren C Teffeau, Carrie M Pirmann, and OCLC Research. 2009. *Scholarly Information Practices in the Online Environment: Themes from the Literature and Implications for Library Service Development*. Dublin, Ohio: OCLC Research. https://www.oclc.org/content/dam/research/publications/library/2009/2009-02.pdf (October 5, 2020).

Pettigrew, Karen E., Raya Fidel, Harry Bruce, and Martha E. Williams. 2001. ‘Conceptual Frameworks in Information Behavior’. In *Annual Review of Information Science and Technology*, Medford: Information Today, 43–78.

Pidgeon, Nick, and Karen Henwood. 2004. ‘Grounded Theory’. In *Handbook of Data Analysis*, eds. Melissa Hardy and Alan Bryman. London: Sage, 625–48. http://methods.sagepub.com/book/handbook-of-data-analysis/n28.xml (October 5, 2020).

Pilat, Dirk, and Yukiko Fukasaku. 2007. ‘OECD Principles and Guidelines for Access to Research Data from Public Funding’. *Data Science Journal* 6: OD4–11.

Pontis, Sheila et al. 2017. ‘Keeping up to Date: An Academic Researcher’s Information Journey’. *Journal of the Association for Information Science and Technology* 68(1): 22–35.

Powell, Ronald R., and Lynn Silipigni Connaway. 2004. *Basic Research Methods for Librarians*. 4th ed. Westport, Conn.: Libraries Unlimited.

Punch, Keith F. 2013. *Introduction to Social Research: Quantitative and Qualitative Approaches*. 3rd edition. London: Sage.

Quandt, Markus, and Reiner Mauer. 2012. ‘Sozialwissenschaften’. In *Langzeitarchivierung von Forschungsdaten. Eine Bestandsaufnahme*, eds. Heike Neuroth et al. Boizenburg: Verlag Werner Hülsbusch, 61–81.

Quińones, Miguel A., J. Kevin Ford, and Mark S. Teachout. 1995. ‘The Relationship Between Work Experience and Job Performance: A Conceptual and Meta-Analytic Review’. *Personnel Psychology* 48(4): 887–910.

RatSWD. 2017. *Die sozial-, verhaltens- und wirtschaftswissenschaftliche Survey-Landschaft in Deutschland - Empfehlungen des RatSWD*. Berlin: Rat für Sozial- und Wirtschaftsdaten (RatSWD). https://www.ratswd.de/dl/RatSWD_Output6_BerichtPanelsurveys.pdf (October 5, 2020).

Recker, Astrid, and Stefan Müller. 2015. 'Preserving the Essence: Identifying the Significant Properties of Social Science Research Data'. *New Review of Information Networking* 20(1–2): 229–35.

Rivera, Gibran, and Andrew Cox. 2014. 'An Evaluation of the Practice Based Approach to Understanding the Adoption and Use of Information Systems'. *Journal of Documentation* 70(5): 878–901.

Robbin, Alice. 1995. 'SIPP ACCESS, an Information System for Complex Data: A Case Study in Creating a Collaboratory for the Social Sciences'. *Internet Research* 5(2): 37–66.

van de Sandt, Stephanie, Sünje Dallmeier-Tiessen, Artemis Lavasa, and Vivien Petras. 2019. 'The Definition of Reuse'. *Data Science Journal* 18(2): 2.

Saracevic, Tefko. 2009. 'Information Science'. In *Encyclopedia of Library and Information Sciences*, eds. Marcia J. Bates and Mary Niles Maack. New York: Taylor & Francis, 2570–86.

Satish, N. G. 1994. 17 *Attitude towards Information*. New Delhi: Concept.

Savolainen, Reijo. 2007. 'Information Behavior and Information Practice: Reviewing the "Umbrella Concepts" of Information-Seeking Studies'. *The Library Quarterly* 77(2): 109–32.

———. 2009. 'Small World and Information Grounds as Contexts of Information Seeking and Sharing'. *Library & Information Science Research* 31(1): 38–45.

———. 2016. 'Elaborating the Conceptual Space of Information-Seeking Phenomena'. *Information Research* 21(3).

———. 2017. 'Information Need as Trigger and Driver of Information Seeking: A Conceptual Analysis'. *Aslib Journal of Information Management* 69(1): 2–21.

———. 2018. 'Information-Seeking Processes as Temporal Developments: Comparison of Stage-Based and Cyclic Approaches'. *Journal of the Association for Information Science and Technology* 69(6): 787–97.

Scheibe, Katrin, Kaja J. Fietkiewicz, and Wolfgang G. Stock. 2016. 'Information Behavior on Social Live Streaming Services'. *Journal of Information Science Theory and Practice* 4(2): 6–20.

Scheuch, Erwin K. 2003. 'History and Visions in the Development of Data Services for the Social Sciences'. *International Social Science Journal* 55(177): 385–99.

Schnell, Rainer, Paul B. Hill, and Elke Esser. 2013. *Methoden Der Empirischen Sozialforschung*. 10., überarbeitete Auflage. München: Oldenbourg.

Shah, Chirag. 2014. 'Collaborative Information Seeking'. *Journal of the Association for Information Science and Technology* 65(2): 215–36.

Sieber, Joan E. 1991. 'Introduction'. In *Sharing Social Science Data. Advantages and Challenges*, Sage focus editions, ed. Joan E. Sieber. Newbury Park: Sage, 1–18.

Slater, Margaret. 1988. 'Social Scientists' Information Needs in the 1980s'. *Journal of Documentation* 44(3): 226–37.

Spink, Amanda, and Jannica Heinström, eds. 2011. *New Directions in Information Behaviour*. Bingley: Emerald.

StataCorp. 2017. *Stata Statistical Software. Release 15*. College Station: StataCorp LLC.

Statistisches Bundesamt. 2018. 'Zahl der Habilitationen 2017 gegenüber Vorjahr geringfügig um 0,3 % gestiegen'. *Statistisches Bundesamt*. https://www.destatis.de/DE/Presse/Pressemitteilungen/2018/07/PD18_242_213.html (October 5, 2020).

Stewart, David W., and Michael A. Kamins. 1993. *Secondary Research: Information Sources and Methods*. 2nd edition. Sage.

Stoan, Stephen K. 1991. 'Research and Information Retrieval Among Academic Researchers: Implications for Library Instruction'. *Library Trends* 39(3): 238–58.

Sun, Guangyuan, and Christopher Soo Guan Khoo. 2017. 'Social Science Research Data Curation: Issues of Reuse'. *Libellarium* 9(2): 59–80.

Tabak, Edin. 2014. 'Jumping between Context and Users: A Difficulty in Tracing Information Practices'. *Journal of the Association for Information Science and Technology* 65(11): 2223–32.

Talja, Sanna. 2005. 'The Domain Analytic Approach to Scholars' Information Practices'. In *Theories of Information Behavior*, ASIST monograph series, eds. Karen Fisher, Sanda Erdelez, and Lynne McKechnie. Medford: Information Today, 123–27.

Talja, Sanna, and Jenna Hartel. 2007. 'Revisiting the User-Centred Turn in Information Science Research: An Intellectual History Perspective'. *Information Research* 12(4). http://www.informationr.net/ir/12-4/colis04.html (October 5, 2020).

Talja, Sanna, and Hanni Maula. 2003. 'Reasons for the Use and Non-use of Electronic Journals and Databases'. *Journal of Documentation* 59(6): 673–91.

Talja, Sanna, and Pamela J. McKenzie. 2007. 'Editors' Introduction: Special Issue on Discursive Approaches to Information Seeking in Context'. *The Library Quarterly* 77(2): 97–108.

Talja, Sanna, Kimmo Tuominen, and Reijo Savolainen. 2005. '"Isms" in Information Science: Constructivism, Collectivism and Constructionism'. *Journal of Documentation* 61(1): 79–101.

Teddlie, Charles, and Abbas Tashakkori. 2009. *Foundations of Mixed Methods Research*. Los Angeles, Calif.: Sage.

Tenopir, Carol et al. 2010. 'Cross Country Comparison of Scholarly E-Reading Patterns in Australia, Finland, and the United States'. *Australian Academic & Research Libraries* 41(1): 26–41.

———. 2011. 'Data Sharing by Scientists: Practices and Perceptions'. *PLOS ONE* 6(6). https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0021101 (October 5, 2020).

Tenopir, Carol, and Donald W. King. 2008. 'Electronic Journals and Changes in Scholarly Article Seeking and Reading Patterns'. *D-Lib Magazine* 14(11/12).

Tenopir, Carol, Donald W. King, Sheri Edwards, and Lei Wu. 2009. 'Electronic Journals and Changes in Scholarly Article Seeking and Reading Patterns'. *Aslib Proceedings* 61(1): 5–32.

Tenopir, Carol, Donald W. King, Jesse Spencer, and Lei Wu. 2009. 'Variations in Article Seeking and Reading Patterns of Academics: What Makes a Difference?' *Library & Information Science Research* 31(3): 139–48.

Toepoel, Vera. 2016. *Doing Surveys Online*. Los Angeles: SAGE.

Torrance, Harry. 2012. 'Triangulation, Respondent Validation, and Democratic Participation in Mixed Methods Research'. *Journal of Mixed Methods Research* 6(2): 111–23.

University of Ljubljana. 2018. *1KA OneKlick Survey. Version 18.10.27*. Ljubljana: University of Ljubljana.

Vakkari, Pertti, and Sanna Talja. 2006. 'Searching for Electronic Journal Articles to Support Academic Tasks. A Case Study of the Use of the Finnish National Electronic Library (FinELib)'. *Information Research* 12(1). https://eric.ed.gov/?id=EJ1104693 (October 5, 2020).

Vardigan, Mary, Pascal Heus, and Wendy Thomas. 2008. 'Data Documentation Initiative: Toward a Standard for the Social Sciences'. *International Journal of Digital Curation* 3(1): 107–13.

Vardigan, Mary, and Cole Whiteman. 2007. 'ICPSR Meets OAIS: Applying the OAIS Reference Model to the Social Science Archive Context'. *Archival Science* 7(1): 73–87.

Weber, Nicholas M. 2013. 'The Relevance of Research Data Sharing and Reuse Studies'. *Bulletin of the American Society for Information Science and Technology* 39(6): 23–26.

Weigl, D. M., K. R. Page, P. Organisciak, and J. S. Downie. 2017. 'Information-Seeking in Large-Scale Digital Libraries: Strategies for Scholarly Workset Creation'. In *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, , 1–4.

Wersig, G., and G. Windel. 1985. 'Information Science Needs a Theory of "Information Actions"'. *Social Science Information Studies* 5(1): 11–23.

White, Howard D. 2011. 'Scientific and Scholarly Network Analysis'. In *The SAGE Handbook of Social Network Analysis*, eds. John Scott and Peter J. Carrington. London: Sage, 271–85.

Widén-Wulff, Gunilla. 2007. *The Challenges of Knowledge Sharing in Practice: A Social Approach*. Oxford: Chandos.

Wijetunge, Pradeepa. 2018. 'Uncertainty Encountered by the Humanities and Social Science Undergraduates in Their Information Seeking Behaviour'. In *Re-Engineering Libraries to Align with Transitioning Educational & Technological Paradigms*, Sri Lanka: Library Network, Eastern University, Sri Lanka. http://archive.cmb.ac.lk:8080/research/handle/70130/4580 (October 5, 2020).

Wilson, Tom D. 1981. 'On User Studies and Information Needs'. *Journal of Documentation* 37(1): 3–15.

———. 1999. 'Models in Information Behaviour Research'. *Journal of Documentation* 55(3): 249–70.

———. 2000. 'Human Information Behavior'. *Informing Science* 3(2): 49–55.

———. 2005. ‘Evolution in Information Behavior Modeling: Wilson’s Model’. In *Theories of Information Behavior*, ASIST monograph series, eds. Karen Fisher, Sanda Erdelez, and Lynne McKechnie. Medford: Information Today, 31–36.

———. 2006. ‘Revisiting User Studies and Information Needs’. *Journal of Documentation* 62(6): 680–84.

———. 2009. ‘Book Review: Everyday Information Practices: A Social Phenomenological Perspective’. *Information Research*. http://informationr.net/ir/reviews/revs327.html (October 5, 2020).

Wissenschaftsrat. 2011. ‘Übergreifende Empfehlungen Zu Informationsinfrastrukturen (Drs. 10466-11)’. https://www.wissenschaftsrat.de/download/archiv/10466-11.pdf (October 5, 2020).

Yoon, Ayoung. 2017. ‘Role of Communication in Data Reuse’. *Proceedings of the Association for Information Science and Technology* 54(1): 463–71.

Yoon, Ayoung, and Youngseek Kim. 2017. ‘Social Scientists’ Data Reuse Behaviors: Exploring the Roles of Attitudinal Beliefs, Attitudes, Norms, and Data Repositories’. *Library & Information Science Research* 39(3): 224–33.

Zenk-Möltgen, Wolfgang et al. 2018. ‘Factors Influencing the Data Sharing Behavior of Researchers in Sociology and Political Science’. *Journal of Documentation* 74(5): 1053–73.

Zimmerman, Ann. 2007. ‘Not by Metadata Alone: The Use of Diverse Forms of Knowledge to Locate Data for Reuse’. *International Journal on Digital Libraries* 7(1–2): 5–16.

Annex

Annex 1: Interview guide (German)

"Auf der Suche nach Daten: das Informationssuchverhalten der NutzerInnen von Umfragedaten"

Leitfaden zum Interview

Mögliche einführende Fragen:

Seit wann beschäftigen Sie sich mit Nutzerinnenanfragen?

Wie viele Nutzerinnenanfragen bearbeiten Sie pro Tag oder Woche im Schnitt?

Welche Anfragen stellen Nutzerinnen und Nutzer an Sie?

Themengebiete:

Welchen Bildungs- oder Berufshintergrund haben die Nutzerinnen und Nutzer?
Beruf; Bildung; Disziplin; Erfahrung.

Wie gut kennen sich die Nutzerinnen und Nutzer mit Umfragedaten aus? Gibt es auch Interessenten, die sich überhaupt nicht mit Umfragedaten auskennen? Fragen sie auch nach anderen Datenarten?
Alternativen: Statistiken, Auswertungen.

Wie sind die Nutzerinnen und Nutzer auf Ihren Service aufmerksam geworden? Woher kommen die Nutzerinnen und Nutzer? Auf welchem Weg nehmen sie Kontakt auf?
Ausbildung; Hinweise von Kollegen bzw. anderen Nutzerinnen und Nutzern; Informationsquellen; Internetrecherche; deutsche und internationale Nutzerinnen und Nutzer; Kommunikationskanäle.

Wofür verwenden die Nutzerinnen und Nutzer die Daten? Welche Aufgaben und Ziele stehen hinter der Nutzung?
Forschungsfragen beantworten; Operationalisierungsfragen; Theorie vorantreiben; Methoden vorantreiben; Replikation; Lehre.

Worauf kommt es Nutzerinnen und Nutzern in Bezug auf die Daten an? Welche Kriterien legen sie an, welche Anforderungen stellen sie an die Daten?
Thematische Relevanz; methodologische Anforderungen; Datenqualität.

Gibt es bestimmte Themen im Sinne von Forschungstrends, die aus den Anfragen erkennbar sind?
z.B. Erforschung von Wandel; Methoden.

Inwiefern nutzen die Nutzerinnen und Nutzer die vorhandene Dokumentation?
Datenkatalog, Codebooks, Informationen auf der Webseite.

Welche Hindernisse und Probleme in Bezug auf die Datennutzung äußern die Nutzerinnen und Nutzer? Welche Hindernisse und Probleme kommen immer wieder vor?
Fehlende Daten; rechtliche Probleme; Datenzugang; Datenqualität; Komplexität; Vergleichbarkeit; persönliche Einschränkungen/ Fähigkeiten/ Infrastruktur.

Entsprach unser Gespräch Ihren Erwartungen?

Annex 2: Informed Consent Form (German)

"Auf der Suche nach Daten: das Informationssuchverhalten der NutzerInnen von Umfragedaten"

Information und Einwilligung
zum Interview und zu den Interviewdaten

[Page 1 of 2]

Projektverantwortliche:

Tanja Friedrich
GESIS – Leibniz-Institut für Sozialwissenschaften
Unter Sachsenhausen 6-8
50667 Köln
Telefon: 0221 47694-457
E-Mail: tanja.friedrich@gesis.org

Projektbeschreibung

In meinem Forschungsprojekt „Auf der Suche nach Daten: das Informationssuchverhalten der NutzerInnen von Umfragedaten" will ich mehr über die NutzerInnen von Umfragedaten erfahren. In diesem Zusammenhang interessieren mich besonders die Anfragen, die diese Personen an Datenarchive und andere datenhaltende Institutionen richten. Um mir ein besseres Bild von diesen Anfragen machen zu können, führe ich mit MitarbeiterInnen dieser Einrichtungen Gespräche zu ihren Erfahrungen mit Beratungsgesprächen. Auch mit Ihnen möchte ich gerne ein Interview zu diesem Thema führen.

Es handelt sich bei diesem Projekt um ein Promotionsvorhaben an der Humboldt-Universität zu Berlin. Die Dissertation wird betreut von Prof. Vivien Petras, PhD, Professorin für *Information Retrieval* am Institut für Bibliotheks- und Informationswissenschaft. Auf Grundlage des Projektes sollen neben der Doktorarbeit auch weitere wissenschaftliche Veröffentlichungen zum Thema entstehen.

Informationen zur Erhebung, Verarbeitung und Archivierung der Interviewdaten

Um die Informationen aus unserem Gespräch auswerten und verarbeiten zu können, werde ich das Interview mit einem Aufnahmegerät und einem Mobiltelefon aufzeichnen. Die

Aufzeichnung werde ich danach in Schriftform bringen. Für die weitere Verarbeitung werde ich Klarnamen von Personen, Projekten oder Institutionen aus den Texten entfernen. Die auf diese Weise verschriftlichten und bearbeiteten Interviews werden GutachterInnen im Rahmen des Promotionsverfahrens zur Einsicht zur Verfügung gestellt. Persönliche Kontaktdaten werden getrennt von den Interviewtexten aufbewahrt und nicht weitergegeben. In den wissenschaftlichen Veröffentlichungen werden immer nur Ausschnitte aus den Interviews zitiert, um sicherzustellen, dass Dritte aus dem dargestellten Gesamtzusammenhang von Ereignissen nicht auf die Identität der interviewten Person schließen können.

Nach Abschluss des Forschungsprojekts sollen die Interviewdaten auch anderen ForscherInnen für ihre wissenschaftliche Arbeit zur Verfügung stehen. Hierzu werden die Interviewtexte in anonymisierter Form an das Datenarchiv des GESIS Leibniz-Instituts für Sozialwissenschaften übergeben. Alle Angaben, die zu einer Identifizierung der befragten Person führen können, werden dazu verändert oder aus dem Text entfernt. Die Tonaufzeichnungen und persönliche Kontaktinformationen werden nach dem Abschluss des Forschungsprojekts nicht archiviert, sondern gelöscht.

[Page 2 of 2]

Einwilligungserklärung zum Interview

Forschungsprojekt: „Auf der Suche nach Daten: das Informationsverhalten der NutzerInnen von Umfragedaten“

Interviewerin: Tanja Friedrich

TeilnehmerIn: _____

Interviewdatum: _____

Ich bin damit einverstanden, an einem Interview im Rahmen des genannten Forschungsprojekts teilzunehmen. Über Inhalte und Ziele des Forschungsprojekts wurde ich informiert. Ich willige ein, dass das Gespräch mit einem Aufnahmegerät und einem Mobiltelefon aufgezeichnet und nach dem Interview durch die Interviewerin verschriftlicht wird. Mit der Weitergabe des schriftlichen Interviewtextes an GutachterInnen im Rahmen des Promotionsverfahrens bin ich einverstanden. Außerdem bin ich einverstanden, dass der Text in anonymisierter Form nach Abschluss des Projektes zur Archivierung an das Datenarchiv von GESIS weitergegeben und durch GESIS anderen ForscherInnen zur

Verfügung gestellt wird. Mir wurde versichert, dass sämtliche Tonaufzeichnungen und persönlichen Kontaktdaten nach Abschluss des Projektes gelöscht werden.

Die Teilnahme an diesem Interview ist freiwillig und ich habe zu jeder Zeit die Möglichkeit, das Gespräch abzubrechen und mein Einverständnis zur Aufzeichnung und Niederschrift des Interviews zurückzunehmen, ohne dass mir daraus irgendwelche Nachteile entstehen.

Ort, Datum, Unterschrift TeilnehmerIn

Ort, Datum, Unterschrift Interviewerin

Annex 3: Interview Guide with Notes (German)

"Auf der Suche nach Daten: das Informationssuchverhalten der NutzerInnen von Umfragedaten"

Leitfaden zum Interview

Mögliche einführende Fragen:

Seit wann beschäftigen Sie sich mit Nutzerinnenanfragen?

Wie viele Nutzerinnenanfragen bearbeiten Sie pro Tag oder Woche im Schnitt?

Welche Anfragen stellen Nutzerinnen und Nutzer an Sie?

Themengebiete:

Welchen Bildungs- oder Berufshintergrund haben die Nutzerinnen und Nutzer?
Beruf; Bildung; Disziplin; Erfahrung.

Wie gut kennen sich die Nutzerinnen und Nutzer mit Umfragedaten aus? Gibt es auch Interessenten, die sich überhaupt nicht mit Umfragedaten auskennen? Fragen sie auch nach anderen Datenarten?
Alternativen: Statistiken, Auswertungen.

Wie sind die Nutzerinnen und Nutzer auf Ihren Service aufmerksam geworden? Woher kommen die Nutzerinnen und Nutzer? Auf welchem Weg nehmen sie Kontakt auf?
Ausbildung; Hinweise von Kollegen bzw. anderen Nutzerinnen und Nutzern; Informationsquellen; Internetrecherche; deutsche und internationale Nutzerinnen und Nutzer; Kommunikationskanäle.

Wofür verwenden die Nutzerinnen und Nutzer die Daten? Welche Aufgaben und Ziele stehen hinter der Nutzung?
Forschungsfragen beantworten; Operationalisierungsfragen; Theorie vorantreiben; Methoden vorantreiben; Replikation; Lehre.

Worauf kommt es Nutzerinnen und Nutzern in Bezug auf die Daten an? Welche Kriterien legen sie an, welche Anforderungen stellen sie an die Daten?
Thematische Relevanz; methodologische Anforderungen; Datenqualität.

Gibt es bestimmte Themen im Sinne von Forschungstrends, die aus den Anfragen erkennbar sind?
z.B. Erforschung von Wandel; Methoden.

Inwiefern nutzen die Nutzerinnen und Nutzer die vorhandene Dokumentation?
Datenkatalog, Codebooks, Informationen auf der Webseite.

Welche Hindernisse und Probleme in Bezug auf die Datennutzung äußern die Nutzerinnen und Nutzer? Welche Hindernisse und Probleme kommen immer wieder vor?
Fehlende Daten; rechtliche Probleme; Datenzugang; Datenqualität; Komplexität; Vergleichbarkeit; persönliche Einschränkungen/ Fähigkeiten/ Infrastruktur.

Entsprach unser Gespräch Deinen Erwartungen?

Annex 4: Initial Codes

Date/Time: 2018-05-11 13:19:31

accessing data depending on legal aspects
acquiring technical skills
adjusting research question with available data
adopting data to create something new
analysing change
applying for data access
appreciating good documentation
appreciating variety of topics
asking data producer for permission to use data
asking data service for advice
asking data service instead of consulting documentation
asking data service instead of consulting website
asking for facts
asking for results
asking methodological questions
asking questions about documentation
asking questions about measured concepts
asking questions about sampling
asking questions about weighting
asking simple questions
asking very particular questions on specific datasets
assessing data based on different criteria
being a business school student
being a dedicated researcher
being a journalist
being a professor
being a teacher
being a university teacher
being academic
being assisted by data service
being clear in requests
being computer literate
being confronted with data early in education
being contacted by data service
being critical about measurements
being deeply immersed in a particular dataset
being denied data access
being experienced in data use
being expert on a particular dataset
being guided by data service
being influenced by educational standards

being information literate
being interested in abstraction
being interested in data quality
being interested in descriptive information
being interested in good documentation
being interested in methods
being oblivious to methodological problems
being offered results
being phd student
being pragmatic
being producer as well as secondary user of data
being referred to alternative datasets
being referred to data service
being referred to documentation
being referred to experts
being referred to information sources
being referred to primary researchers
being referred to special services
being referred to the library
being referred to training
being researcher
being senior researcher
being seriously interested in data with restricted access
being unclear in requests
being undergraduate
being unexperienced in data use
being unfamiliar with data formats
being unskilled in empirical methodology
being unskilled in survey data use
being vague in requests
being very experienced in data use
being very specific in requests
belonging to commercial consulting companies
belonging to governmental institutions
belonging to non-university research institutes
belonging to private research institutes
building hypotheses
calling data service
choosing between alternative datasets
choosing between alternative datasets based on quality of documentation
combining methodological with substantial intersts
coming from different disciplines
coming from disciplines other than social sciences
confusing data with other types of information
consulting data service
consulting data service repeatedly
consulting questionnaires

consulting with data service on skills
declaring intended usage
depending on national research agendas
depending on schedules
designing own research
designing questionnaires
detecting errors in data
developing own questions
developing research questions based on available data
doing secondary analysis
downloading datasets
downloading documentation
downloading questionnaires
e-mailing data service
enquiring about alternative datasets
enquiring about response rates
enquiring costs of data use
enquiring data access
enquiring details of particular datasets
enquiring in very particular details of datasets
enquiring inconsistencies in datasets
enquiring questionnaire wordings
evaluating available data with regard to research questions
exemplifying methodological aspects in answering research questions
expecting data service to do statistical analyses
expecting data to comply with certain standards
exploring data service websites
facing barriers in data access
facing barriers related to infrastructure
facing complex responsibilities
facing complexity of datasets
facing detailed documentation
facing discontinuity in longitudinal data
facing extensive data collections
facing high quality in methodology
facing inadequate information
facing language barriers
facing legal barriers in data analysis
facing legal barriers in data gathering
facing limited documentation
facing low response rates
facing misleading documentation
facing multiple sources of documentation
facing problems of comparability
facing problems of representativity
facing problems of sampling
facing problems with website information

facing retrieval problems
facing technical barriers
facing too much information
facing well-structured datasets
finding data through search engines
finding popular datasets
finding that available data do not fit research question
finding that needed data do not exist
finding their way to experts
focusing own work on one particular survey programme
following research trends
gaining reputation
GOAL OF SEEKING
goal of seeking: analysing a specific concept
goal of seeking: analysing a specific population
goal of seeking: apply specific methods
goal of seeking: doing secondary analysis
goal of seeking: finding indicators to measure a specific concept
goal of seeking: finding specific measurements
goal of seeking: proving an assumption
goal of seeking: researching a specific topic
goal of seeking: telling a story
goal of seeking: testing specific hypotheses
goal of seeking: writing a seminar paper
goal of seeking: writing a thesis
having a non-academic audience
having access to questionnaires
having access to variables
having an understanding of what data are
having clearly defined research questions
having data service check data
having data service check results with regard to legal aspects
having data service convert data into other formats
having data service perfom searches
having data service perform statistical analyses
having data service produce simple statistics
having data service translate information
having datasets delivered
having different educational and occupational background
having free access to data
having knowledge about the background of data collection
having less information than data service
having more knowledge on a particular dataset than data service
having problems with specific datasets
having various requests
having varying experience with survey data
having varying knowledge on a specific dataset

having varying skills and compentences
identifying data from literature (chaining)
identifying literature from data documentation (backward chaining)
indicating relevant literature
investing in originality
knowing about data catalogues
knowing about data service
knowing about datasets from media
knowing about popular datasets
lacking experience with complex datasets
learning about alternative datasets
learning about data access
learning about data from literature
learning about data from media
learning about data from other sources
learning about data from professors
learning about data service
learning about free access to data
learning about popular datasets
learning about possibilites of online analysis
linking data
looking for data
looking for data at a young age
looking for data from large survey programmes
looking for data in an early stage of research
looking for data internationally
looking for information in general
looking for information on specific datasets
looking for international survey programmes
looking for known data
looking for literature
looking for longitudinal data
looking for panel data
looking for questionnaires
looking for questions
looking for recent data
looking for representative data
looking for results
looking for specific datasets
looking for statistical evaluations
making an effort to work with data
making general enquiries on data
making mistakes in data use
making own calculations
making relevance judgements
needing documentation
needing help of intermediaries to access data

needing help of intermediaries to find data
needing more detailed documentation
needing more support
needing non-proprietary data formats
needing simpler information
not appreciating available documentation
not asking basic questions
not being academic
not being interested in data quality
not being interested in explanation
not being seriously interested in the data
not being trained in legal aspects
not citing data properly
not considering documentation
not consulting data service
not enquiring about alternative datasets
not enquiring about response rates
not having an understanding of what data are
not having knowledge about the background of data collection
not knowing about data catalogues
not knowing about legal aspects of data use
not knowing how to apply weighting
not learning about unpopular datasets
not replicating research from others
not scrutinizing the data
not understanding documentation
not understanding terms of data access
not using DOIs
not using unpopular datasets
not working purely methodological
not working purely theoretical
not working with scientific methods
not working with statistical software
ordering datasets
ordering datasets using a form
passing on datasets
paying for data access
performing in-depth research with a single dataset
performing simple statistics
performing statistical analyses
putting quality over price
putting topical relevance over quality
reading data newsletters
reading data wrong
reanalysing details
receiving advice on conceptual measures
receiving advice on crucial issues

receiving advice on weighting
receiving datasets from others
receiving detailed information on sample sizes
receiving information on datasets via mailing lists
receiving training for specific datasets
recommending data service
registering with a data portal or catalogue
replicating research from others
requesting data based on citations
requesting data for academic research
requesting data for research
requesting data for teaching
requesting data from other countries
requesting data from specific studies
requesting data on specific topics
requesting data with restricted access
requesting other data than survey data
requesting ready-made calculations
requesting specific datasets
requesting specific variables
resolving DOIs
reusing questions
scrutinizing data and questions
searching datasets
searching question texts
searching variables
seeking advice in statistical programming
seeking help of data experts
seeking technical advice
struggling with complexity of datasets
subscribing to a mailing list
using a data catalogue
using an omnibus survey
using available indicators to measure another concept
using data for academic research
using data for publications
using data for research
using data for statistics training
using data for teaching
using data for theses
using data for varying purposes
using data from large survey programmes
using data portal
using data that are freely available
using discipline specific jargon
using documentation
using DOIs

using English documentation instead of German documentation
using high quality datasets
using low quality datasets
using monopolistic studies
using panel data
using popular datasets
using popular methods
using relatively unknown datasets
visiting data service
working comparatively
working cooperatively
working empirically
working in large scale research projects
working less theoretically
working on a theoretical basis
working on false assumptions
working on popular topics
working with complex datasets
working with data as part of education
working with data repeatedly
working with literature
working with longitudinal data
working with methodological rigour
working with sensitive data
working with statistical software

Annex 5: Initial Code Families

Date/Time: 2018-05-11 13:21:09

Kodefamilie: Being diversely skilled

Erstellt: 2016-08-02 10:39:57 (Super)

Kodes (21): [acquiring technical skills] [being computer literate] [being critical about measurements] [being experienced in data use] [being information literate] [being oblivious to methodological problems] [being unexperienced in data use] [being unfamiliar with data formats] [being unskilled in empirical methodology] [being unskilled in survey data use] [being very experienced in data use] [confusing data with other types of information] [having an understanding of what data are] [having varying experience with survey data] [having varying skills and compentences] [lacking experience with complex datasets] [not being trained in legal aspects] [not having an understanding of what data are] [not knowing about legal aspects of data use] [not knowing how to apply weighting] [struggling with complexity of datasets]

Zitat(e): 47

Kodefamilie: Being influenced by external factors

Erstellt: 2016-08-02 10:22:15 (Super)

Kodes (4): [accessing data depending on legal aspects] [being influenced by educational standards] [depending on national research agendas] [depending on schedules]

Zitat(e): 16

Kodefamilie: Employing different styles of request

Erstellt: 2016-08-02 10:36:42 (Super)

Kodes (7): [being clear in requests] [being unclear in requests] [being vague in requests] [being very specific in requests] [having various requests] [looking for data] [using discipline specific jargon]

Looking for data

Zitat(e): 8

Kodefamilie: Facing problems and barriers

Erstellt: 2016-08-02 11:49:59 (Super)

Kodes (18): [being denied data access] [facing barriers in data access] [facing barriers related to infrastructure] [facing complex responsibilities] [facing complexity of datasets] [facing discontinuity in longitudinal data] [facing inadequate information] [facing language barriers] [facing legal barriers in data analysis] [facing limited documentation] [facing misleading documentation] [facing multiple sources of documentation] [facing problems with website information] [facing retrieval problems] [facing technical barriers] [facing too much information] [finding that available data do not fit research question] [finding that needed data do not exist]

Zitat(e): 43

Kodefamilie: Having a certain affiliation, profession, or education

Erstellt: 2016-08-02 10:25:13 (Super)

Kodes (19): [being a business school student] [being a journalist] [being a professor] [being a teacher] [being a university teacher] [being academic] [being phd student] [being researcher] [being senior researcher] [being undergraduate] [belonging to commercial consulting companies] [belonging to governmental institutions] [belonging to non-university research institutes] [belonging to private research institutes] [coming from different disciplines] [coming from disciplines other than social sciences] [having different educational and occupational background] [not being academic] [working in large scale research projects]

Zitat(e): 56

Kodefamilie: Interacting with data service

Erstellt: 2016-08-02 10:26:59 (Super)

Kodes (71): [applying for data access] [asking data service for advice] [asking data service instead of consulting documentation] [asking data service instead of consulting website] [asking for facts] [asking for results] [asking methodological questions] [asking questions about documentation] [asking questions about measured concepts] [asking questions about sampling] [asking questions about weighting] [asking simple questions] [asking very particular questions on specific datasets] [being assisted by data service] [being contacted by data service] [being guided by data service] [being offered results] [being referred to alternative datasets] [being referred to documentation] [being referred to experts] [being referred to information sources] [being referred to primary researchers] [being referred to special services] [being referred to the library] [being referred to training] [calling data service] [consulting data service] [consulting data service repeatedly] [consulting with data service on skills] [declaring intended usage] [e-mailing data service] [enquiring about alternative datasets] [enquiring about response rates] [enquiring costs of data use] [enquiring data access] [enquiring details of particular datasets] [enquiring in very particular details of datasets] [enquiring inconsistencies in datasets] [enquiring questionnaire wordings] [expecting data service to do statistical analyses] [exploring data service websites] [finding their way to experts] [having data service check data] [having data service check results with regard to legal aspects] [having data service convert data into other formats] [having data service perfom searches] [having data service perform statistical analyses] [having data service produce simple statistics] [having data service translate information] [indicating relevant literature] [learning about data access] [making general enquiries on data] [receiving advice on conceptual measures] [receiving advice on crucial issues] [receiving advice on weighting] [requesting data based on citations] [requesting data for academic research] [requesting data for research] [requesting data for teaching] [requesting data from other countries] [requesting data from specific studies] [requesting data on specific topics] [requesting data with restricted access] [requesting other data than survey data] [requesting ready-made calculations] [requesting specific datasets] [requesting specific variables] [seeking advice in statistical programming] [seeking help of data experts] [seeking technical advice] [visiting data service]

Zitat(e): 133

Kodefamilie: Knowing and learning about data

Erstellt: 2016-08-02 11:07:10 (Super)

Kodes (12): [identifying data from literature (chaining)] [identifying literature from data documentation (backward chaining)] [knowing about datasets from media] [knowing about popular datasets] [learning about alternative datasets] [learning about data from literature] [learning about data from media] [learning about data from other sources] [learning about

data from professors] [learning about popular datasets] [not learning about unpopular datasets] [requesting data based on citations]

Zitat(e): 21

Kodefamilie: Satisfying a particular goal

Erstellt: 2016-08-02 11:36:32 (Super)

Comment:

Wilson defines information seeking as the "purposive seeking for information as a consequence of a need to satisfy some goal." (Wilson 2000, 49)

Kodes (12): [goal of seeking: analysing a specific concept] [goal of seeking: analysing a specific population] [goal of seeking: apply specific methods] [goal of seeking: doing secondary analysis] [goal of seeking: finding indicators to measure a specific concept] [goal of seeking: finding specific measurements] [goal of seeking: proving an assumption] [goal of seeking: researching a specific topic] [goal of seeking: telling a story] [goal of seeking: testing specific hypotheses] [goal of seeking: writing a seminar paper] [goal of seeking: writing a thesis]

Zitat(e): 39

Annex 6: Focused Codes

Date/Time: 2018-05-14 11:51:35

BACKGROUND
BACKGROUND_being an experienced researcher
BACKGROUND_being influenced by domain specifics
BACKGROUND_coming from another discipline
BACKGROUND_having non-academic background
BACKGROUND_having varying educational and professional backgrounds
BACKGROUND_using data in journalism
BACKGROUND_working empirically
BACKGROUND_working for governmental institutions
BACKGROUND_working theory based
BARRIERS
BARRIERS_DATA_facing content that is not self-explanatory
BARRIERS_DATA_facing data of varying scientific quality
BARRIERS_DATA_facing different circulating versions of datasets
BARRIERS_DATA_facing errors in data
BARRIERS_DATA_facing incongruous data from different studies
BARRIERS_DATA_facing language barriers
BARRIERS_DATA_facing low response rates
BARRIERS_DATA_facing sampling issues
BARRIERS_DATA_requiring data that do not exist
BARRIERS_DOCUMENTATION_being overchallenged by exhaustive documentation
BARRIERS_DOCUMENTATION_facing misleading documentation
BARRIERS_DOCUMENTATION_facing restricted documentation
BARRIERS_DOCUMENTATION_facing varying standards
BARRIERS_INFORMATION_being challenged by complex website information
BARRIERS_INFRA_facing limited capacity of data service
BARRIERS_INFRA_lacking computing capacity
BARRIERS_INFRA_not having access to statistical software
BARRIERS_LEGAL_facing barriers regarding commercial use of data
BARRIERS_LEGAL_facing legal barriers in data access
BARRIERS_LEGAL_facing legal barriers in working with data
BARRIERS_LEGAL_revealing personal information and interest when ordering data
BARRIERS_SEEKING_being overchallenged by the research data landscape
BARRIERS_SEEKING_facing failing online services
BARRIERS_SKILL_facing complex datasets
COMMUNITY
COMMUNITY_being a data service power user
COMMUNITY_being confronted with different responsibilities
COMMUNITY_being expert for a certain dataset
COMMUNITY_being producer and user of data
COMMUNITY_being referred to data service
COMMUNITY_being referred to experts on specific datasets
COMMUNITY_being referred to primary investigators

COMMUNITY_benefiting from networking community members
COMMUNITY_benefiting from shared responsibilities
COMMUNITY_cleaving to data from a particular study
COMMUNITY_contributing to documentation
COMMUNITY_detecting and reporting errors
COMMUNITY_employing personal contacts to find or get access to data
COMMUNITY_receiving news on specific datasets
COMMUNITY_receiving training for specific datasets
COMMUNITY_sharing data informally
CONTEXT
CONTEXT_INFRA_having easy access to data
GOAL
GOAL_INTEREST_being inspired to do research
GOAL_INTEREST_doing spatial analysis
GOAL_INTEREST_looking for results rather than data
GOAL_INTEREST_studying change
GOAL_INTEREST_working comparatively
GOAL_INTEREST_working on specific populations
GOAL_INTEREST_working with international survey programmes
GOAL_INTEREST_working with sensitive data
GOAL_METHOD_applying multivariate statistics
GOAL_METHOD_doing georeferencing
GOAL_METHOD_doing multilevel analysis
GOAL_METHOD_doing record linkage
GOAL_METHOD_working with longitudinal data
GOAL_METHOD_working with panel data
GOAL_NEED_needing simple analyses or results
GOAL_SUCCESS_doing original research
GOAL_SUCCESS_following a trend to work with data
GOAL_SUCCESS_following trends in methodology
GOAL_SUCCESS_needing recent data
GOAL_SUCCESS_not being interested in replications
GOAL_SUCCESS_preferring data from prestigious primary investigators
GOAL_SUCCESS_researching remarkable topics
GOAL_SUCCESS_researching trending topics
GOAL_SUCCESS_seeking prestige
GOAL_SUCCESS_seeking to get published
GOAL_SUCCESS_working with high quality data
GOAL_TASK_developing new indices from existing data
GOAL_TASK_intending commercial use of data
GOAL_TASK_learning to work with data
GOAL_TASK_making simple descriptive analyses
GOAL_TASK_measuring concepts of interest
GOAL_TASK_sorting out subject of research at an early stage
GOAL_TASK_teaching data use
GOAL_TASK_using data for methodological exercise
GOAL_TASK_using data for replication

GOAL_TASK_using data for teaching
GOAL_TASK_using data in research projects
GOAL_TASK_using existing data to generate hypotheses
GOAL_TASK_using measures for own data gathering
GOAL_UNCLEAR_collecting datasets
GOAL_UNCLEAR_preferring large survey programmes
GOAL_UNCLEAR_preferring popular datasets
GOAL_UNCLEAR_working with unique datasets
REQUIREMENTS
REQUIREMENTS_being required to work with data from early on in education
REQUIREMENTS_depending on academic requirements
REQUIREMENTS_developing research question from available data
REQUIREMENTS_needing data or analysis as quickly as possible
REQUIREMENTS_requesting recommended or stipulated datasets
REQUIREMENTS_requiring access to sensitive data
REQUIREMENTS_requiring data that fit methodological approach
REQUIREMENTS_requiring data that fit research question
REQUIREMENTS_requiring detailed documentation
REQUIREMENTS_requiring flawless data
REQUIREMENTS_undervaluing data quality
SEEKING
SEEKING_CITATIONS_using DOI citations
SEEKING_CITATIONS_using frequently cited data
SEEKING_DOCUMENTATION_facing comprehensive documentation
SEEKING_DOCUMENTATION_making use of documentation
SEEKING_RELEVANCE_adjusting research questions with available data
SEEKING_RELEVANCE_disregarding information on data quality
SEEKING_RELEVANCE_finding that available data do not fit research question
SEEKING_RELEVANCE_performing simple analyses to judge relevance
SEEKING_RELEVANCE_performing simple quality checks on data
SEEKING_SEARCHING_formulating conceptual queries
SEEKING_SEARCHING_scanning datasets for relevance
SEEKING_SEARCHING_searching known datasets
SEEKING_SEARCHING_searching variables
SEEKING_SOURCE_consulting intermediaries
SEEKING_SOURCE_having a choice of high quality data collections
SEEKING_SOURCE_having available a diversity of information channels
SEEKING_SOURCE_having students perform searches
SEEKING_SOURCE_learning about data from literature
SEEKING_SOURCE_learning about data from the media
SEEKING_SOURCE_learning about data in academic or educational contexts
SEEKING_SOURCE_learning about data service
SEEKING_SOURCE_looking for and using data from other disciplines
SEEKING_SOURCE_looking for known data
SEEKING_SOURCE_performing web searches to find data
SEEKING_SOURCE_receiving biased data advertisement
SEEKING_SOURCE_using data search engines

SEEKING_SOURCE_using data service repeatedly
SEEKING_SOURCE_using omnibus surveys
SKILL
SKILL_employing simple statistics
SKILL_NEG_being oblivious to documentation
SKILL_NEG_being oblivious to errors in data
SKILL_NEG_being oblivious to methodological restrictions
SKILL_NEG_lacking knowledge in statistical software
SKILL_NEG_lacking knowledge in statistics
SKILL_NEG_lacking legal knowledge
SKILL_NEG_making mistakes in reading data
SKILL_NEG_making poor analyses
SKILL_NEG_not being (survey) data literate
SKILL_NEG_not being able to understand documentation
SKILL_NEG_not knowing how to apply weighting
SKILL_NEG_not knowing how to obtain data
SKILL_POS_being skilled in finding data
SKILL_POS_detecting errors in datasets
SKILL_POS_having empirical and statistical skills
SKILL_working with noncomplex datasets
SUPPORT
SUPPORT_ANALYSIS_being offered help with analysis
SUPPORT_ANALYSIS_being offered results of simple analyses
SUPPORT_ANALYSIS_having students perform simple analyses
SUPPORT_ANALYSIS_needing help with data analysis
SUPPORT_ANALYSIS_needing help with weighting
SUPPORT_being offered additional information by experts
SUPPORT_being referred to commercial data services
SUPPORT_being referred to documentation
SUPPORT_being referred to literature on the data
SUPPORT_being referred to website information
SUPPORT_DATASET_being offered alternative data
SUPPORT_DATASET_being offered pre-releases of data
SUPPORT_DATASET_being offered specifically processed data
SUPPORT_DATASET_being offered useful tools to work with data
SUPPORT_DATASET_benefiting from expert knowledge
SUPPORT_DATASET_needing help in understanding details of data
SUPPORT_DATASET_needing help to access data
SUPPORT_DATASET_needing help to work with data
SUPPORT_DOCUMENTATION_needing help with documentation
SUPPORT_DOCUMENTATION_receiving information through personal requests instead of documentation
SUPPORT_METHODS_needing advice on methodology
SUPPORT_METHODS_receiving training to work with data
SUPPORT_PERSONAL_being updated on problems with an ordered dataset
SUPPORT_PERSONAL_preferring personal contact over website information
SUPPORT_PERSONAL_preferring written over verbal requests

SUPPORT_receiving recommendations on training
SUPPORT_receiving taylored services according to skill
SUPPORT_RESEARCH_needing help with research design
SUPPORT_saving time by using data service
SUPPORT_SEEKING_being introduced to data search tools
SUPPORT_SEEKING_depending on data service when looking for data
SUPPORT_SEEKING_needing professional help to find data
SUPPORT_TECHNICAL_needing technical data service

Annex 7: Memo "Errors in data or users' mistakes?"

Memos

Date/Time: 2019-12-06 11:18:50

MEMO: Errors in data or users' mistakes? (1 Zitat) (Super, 2016-06-17 09:17:15)
P14: 20160701Tn04.txt:
(227:227)
Keine Kodes
keine Memos
Typ: Commentary

It is natural that users make mistakes when they are working with data. For example, they miss information that is there. It is equally natural, that datasets contain errors, for example missing variables. At the very outset of a data use case, the user cannot know, whether there are any errors in the dataset. This means that encountering errors is always a possibility. However, making mistakes when working with the data is always a possibility, too. So, there is a point in certain usage scenarios, where a user encounters an inconsistency of whatever kind. Possible reasons are either an error in the data or a mistake on the side of the user. Presumably the user will at first check their own procedures for possible mistakes. If they find a mistake, they will correct their path and continue work with the data. If they don't find a mistake, perhaps having taken multiple loops or considered multiple sources of mistakes, they will assume an error in the dataset. What is happening here can be described as activities of verifying as identified and defined by Ellis et al. (1993). What is remarkable though is that there seem to be some cases where users miss rather obvious mistakes that they have made and quickly assume errors in data instead and resort to data service. Several questions are manifesting here: Are there different levels of resilience in users that determine how much effort they invest in checking for possible mistakes made by themselves? Does the fact that there is an approachable data service influence their willingness to check for own mistakes more intensively? Would they keep checking if there was no one to ask whether there was an error in the data? Does the probability of making mistakes correlate with any demographic variables, for example, age, education, field of study? Does the resilience correlate with any of these variables?

Annex 8: Memo "Calling data service instead of using documentation"

Memos

Date/Time: 2019-12-06 11:17:10

MEMO: calling data service instead of using documentation (3 Zitate) (Super, 2016-06-17 14:12:20)
P16: 20160720Tn06.txt:
(75:75), (119:127), (159:187)
Keine Kodes
keine Memos
Typ: Commentary

Some users tend to rather call data service than to inform themselves via documentation or information on websites. This may be for several reasons ...

... on the part of the user: varying information literacy; lacking familiarity with this type of information; time constraints ...

... on the part of the available information: documentation is difficult to understand; too much information to see through; web contents change frequently, sometimes resulting in information loss (broken links) or confusion (scattered information) ...

Maybe some users prefer personal contact over reading information material.

This phenomenon may also be related to the phenomenon of "errors in data or users' mistakes" (see memo) which also leads to increased personal enquiries.

Annex 9: Memo "Dataset communities"

Memos

Edited by: Super

MEMO: Dataset communities (0 Zitate) (Super, 2016-07-08 09:24:27)
Keine Kodes
keine Memos
Typ: Commentary

Large survey programmes produce popular datasets. These datasets are generally created by more than one primary researcher. The primary researchers that are responsible for a large survey programme usually are established experts in their discipline. The datasets, which they produce, are highly visible and receive special treatment in terms of data preparation by respective experts, for example in data archives. Commonly there are more than one of these data curators that are concerned with one survey programme.

Primary researchers and data curators work in close cooperation to create datasets that are attractive to secondary users. From within these activities, a community emerges around the produced datasets. Dataset communities of this kind are made up by primary researchers, data curators, data archivists, data librarians and other data service personnel, sometimes also data collectors and further actors who are concerned with the production, curation, distribution, and archiving of the data. But most notably, these communities also include the secondary users, who are, in most cases, the people that the popular datasets are made for. Sometimes these users are involved in data creation, for example by offering them the opportunity to suggest questions. And sometimes these users apply themselves in data improvement, for example, when they detect and report errors in the data. Sometimes the primary researchers also act as users in this regard; and sometimes they make suggestions for improvement on behalf of the secondary users.

Persons and institutions working for large survey programmes respond to the existence of dataset communities with several supportive structures and mechanisms. For example, they create dataset specific mailing lists; they organise "meet the data" events; they supply teachers with datasets. It would be interesting to find out, who uses these offers and who the very active secondary users are. There are secondary users, who work with one dataset during their entire career; but there should be more people who work with several datasets - there is the whole spectrum of affinity. Does the active dataset community include researchers from the whole spectrum equally? What is active participation anyway? Are there distinctive levels of participation? Is there a correlation between the other detected phenomena - consulting data service instead of documentation and reporting alleged errors?

Annex 10: Memo "Large survey programmes"

Memos

Date/Time: 2019-12-06 11:19:14

MEMO: Large survey programmes (1 Zitat) (Super, 2016-07-08 10:23:39)
P15: 20160707Tn05.txt:
(107:107)
Keine Kodes
keine Memos
Typ: Commentary

In survey research, there are a couple of large survey programmes that are designed for secondary use and thus are funded extensively. Compared to data from smaller surveys, these data usually have extended (added value) documentation and special data services, such as meet the data workshops. Large survey programmes are advertised prominently and often carry a certain prestige that makes many researchers want to work with them.

Large survey programmes are ...
... producing data specifically for secondary use (at least in most cases)
... producing popular datasets
... often producing comparative and longitudinal data
... often including diverse and recent topics
... designed by established researchers
... curated intensively
... kept available for as long as possible (long term archiving)

It seems to me that, compared to literature, the production of data in general and of large datasets in particular requires large investment. The reads of a popular paper may be much higher than the uses of a popular dataset. But the time and resources invested in producing a dataset and in using it seem to compensate for that. Sure enough, articles and datasets serve different purposes. However, we need to find out what makes data use special instead of adopting existing knowledge about information behaviour indiscriminately. For example, a repository for articles might not be suitable to include datasets.

Annex 11: Memo "Concepts and Indicators in secondary analysis"

Memos

Date/Time: 2018-05-14 11:53:21

MEMO: Concepts and indicators in secondary analysis (0 Zitate) (Super, 2016-07-12 10:03:28)
Keine Kodes
keine Memos
Typ: Commentary

People looking for data to do secondary analysis usually intend to investigate their own research questions (except for the few people who want to merely calculate replications). Commonly, social researchers follow a research process where they at first define the variables or concepts that determine their research question (on the difference between variables and concepts see Bernard 2013, 34). When collecting own data, they proceed with the development of measurements and instruments (a questionnaire) to enter the field with. They do this by identifying indicators (or indicants) to measure the concepts of interest and by developing operational definitions of the concepts according to the identified indicators. They formulate their questionnaire accordingly. Measurement design is a crucial step in empirical research and "considerable attention must be given to identifying valid and reliable measures at the onset of the study" (Mueller 2004, 164).

Secondary researchers are not collecting their own data but are looking for available data to measure their concepts with. There are several strategies to find such data. For instance, they can try and find the concept of interest in abstracts or descriptions of data in data catalogues. That way they can find data that were collected by primary researchers with regard to the concept of interest. This way of data retrieval is error-prone for several reasons. For one there might be a different understanding of a concept in different disciplines, for example, in sociology and economics. Different understandings call for different indicators and thus lead to different operational definitions. Differences of this kind can also occur within single fields where researchers follow various schools of thought or metatheoretical approaches. This variation in operational definitions leads the secondary researcher to find data that are somehow related to their concept of interest but not necessarily congruent with their own understanding of the concept.

To a certain extent, this problem of incongruency underlies all secondary research, even within a field, school of thought, or metatheoretical approach. This is because standardisation of measures is indeed restricted to certain key concepts in each discipline. Only in cases where "theory in a specialty area is well established and there already exists a strong research tradition, [...] valid and reliable measures likely already exist" (Mueller 2004, 164). By drawing on standardised measures, reuse with regard to comparability or development over time becomes possible, and accumulation of knowledge in the field becomes more likely (Mueller 2004, 164). In the case of missing established measurements (predominantly in exploratory research, cf. Mueller 2004, 164), the researcher can, however, adopt some common approach in operationalising their concept and make this approach

explicit. Operationalisation is a process of defining concepts by identifying and framing measurable indicators for these concepts. This process is per se individual and the result will never be the exact same operationalisation like any other one made by another researcher.

In secondary research, it is challenging to find data on a concept of interest. Only the simplest concepts are measured directly, with single indicators, e.g. the concept 'age' may be measured by asking a respondent: "How old are you?" (other concepts that can be measured with a single indicator include 'income' and 'education', cf. Mueller 2004, 164). All other concepts such as 'social status', 'political conservatism', 'marital satisfaction', or the like are more complex and therefore measured indirectly, by asking questions on multiple indicators that have been identified by defining the concept (cf. ibd.). Where data on a concept, defined by indicators, cannot be found, researchers have to identify indicators that define their concept of interest by themselves and then go an search for data that are results of the measurement of these indicators. They can then combine these indicators and the associated data to measure the concept. At this point it is important to note, that indicators that have been used to define and measure a certain concept, might as well be helpful to define and measure another concept (see the neighbour politics conversation example). This is why the possibility to search data on the question level is important.

Another aspect which seems to be important in this context is that secondary researchers may be inclined to reuse data on concepts of interest indiscriminately, without checking the operational definition and measurements first. Participants in the interviews have indicated this independently. This practice bears the risk of measurement error, in particular of measurement invalidity, if the applied measure does not capture the targeted concept sufficiently (cf. Mueller 2004, 163).

Definitions:

"Operationalization [is] [t]he translation of an abstract concept (e.g. social status) into something which can be observed (e.g. occupation)." (Miller/Wilson 1983, p. 80)

"An indicator, together with rules for using and interpreting it, is an operational definition of a concept." (Miller/Wilson 1983, p. 21)

"[A] [c]oncept [is] [a] mental construct which selects and summarizes an aspect of the observable world for theoretical attention [...] Concepts as such are non-observable entities, being pure thought constructs, e.g. social mobility, intelligence, suicide, clan, demand, and they must be interpreted by indicators in order to be used empirically." (Miller/Wilson 1983, p. 21)

Literature:

Miller, Patrick McC./ Wilson, Michael J. (1983): A Dictionary of Social Science Methods. Chichester: Wiley.

Mueller, Charles W. (2004): Conceptualization, Operationalization, and Measurement. In: Lewis-Beck, Michael S./ Bryman, Alan/ Liao, Tim Futing (Eds.): The Sage Encyclopedia of Social Science Research Methods. Volume 1. Thousand Oaks: Sage. 161-165.

Annex 12: Memo "People looking for data do what works for them"

Memos

Date/Time: 2019-12-06 11:19:39

MEMO: People looking for data do what works for them (2 Zitate) (Super, 2017-03-03 16:18:37)
P 2: 20160609Tn01.txt:
(403:403)
P 4: 20160616Tn02.txt:
(155:155)
Kodes: [COMMUNITY_being a data service power user] [COMMUNITY_cleaving to data from a particular study] [SKILLS_employing simple statistics] [SKILLS_working with noncomplex datasets]
keine Memos
Typ: Commentary

People who are looking for data or working with data often behave in a way that works for them in terms of reaching their goals. Sometimes they just stick to a strategy or practice that they have acquired and employed successfully in the past. This relates to people or services that they have encountered as well as to methods that they already know. These people rely on experiences that they have made and on the skills that they already have.

Annex 13: Memo "Trust"

Memos

Date/Time: 2019-12-06 11:14:18

MEMO: Trust (1 Zitat) (Super, 2017-03-15 12:53:53)
P15: 20160707Tn05.txt:
(71:79)
Keine Kodes
keine Memos
Typ: Commentary

Good data service includes quality checks of data. These checks may require great expense on the part of the data centre. However, they cannot leave these checks to the researchers because they don't have the necessary information (e.g., raw files; interviewer information) or contacts (e.g., to the field institute) to perform them. Also, it would require a much greater expense, if every researcher were to act on their own. It is one of the core responsibilities of data centres or archives to provide quality checked data along with the necessary documentation. Data centres have to run several routine checks to ensure basic quality of the data that they distribute. This service is at the core of trustworthiness of data centres. There are other players involved that have their own responsibilities regarding data quality, e.g., the interviewers and the field institutes.

Researchers who are looking for data have different channels and sources that they can use. But only data centres and archives give them a foundation of trust with regard to data quality and data integrity. Checking data for quality and integrity is costly but necessary for research that can be trusted and for trust in science in general. There is no rule as to how far these checks should go. But it can be expected that large survey programmes with added value data processing have been checked the most thoroughly. After all, even in large survey programmes not everything can be checked and not every problem of quality or integrity can be outruled. But with more intensive data processing, problems with quality and integrity become less likely.

Annex 14: Memo "Classes of users"

Memos

Date/Time: 2019-12-06 11:17:39

MEMO: Classes of users (2 Zitate) (Super, 2017-03-16 15:02:46)
P15: 20160707Tn05.txt:
(159:163), (247:247)
Keine Kodes
keine Memos
Typ: Commentary

When it comes to the extent of data service, there are different classes of users who receive different intensity of service. For example, one participant (Tn04) indicated that students were not accepted as guest researchers for sensitive data. Another participant (Tn05) said that they refused even simple analyses for a student user due to lack of time. The same participant indicated that, on the other hand, they had performed advanced data processing for a renowned professor but also had turned down another professor's request for such service before.

One particular factor that determines data centres or archives to perform extended service is that they view the more experienced researchers (professors) as multipliers (Tn05) who make advertisement for datasets or data services. In that case, extended service is an investment in visibility.

Another reason (according to Tn05) that motivates data centres and archives to perform advanced data processing is that they assume that the results may be beneficial for other users. It seems to be more likely that requests with such potential come from advanced researchers.

One participant (Tn05) stated that the survey programme that they were responsible for was explicitly targeted to academics as well as a non-academic audience. The programme even offers specific information and tools for journalists or students. However, this was the same participant who indicated earlier that they did not perform simple analyses as a service to unexperienced users. So, while the non-academic audience is important to the programme, they don't receive treatment that is specific to their requests but rather prefabricated information and tools that is believed to match their needs.

Annex 15: Memo "Problem solving by community involvement"

Memos

Date/Time: 2019-12-06 11:20:09

MEMO: Problem solving by community involvement (0 Zitate) (Super, 2017-05-30 15:07:08)
Keine Kodes
keine Memos
Typ: Theory

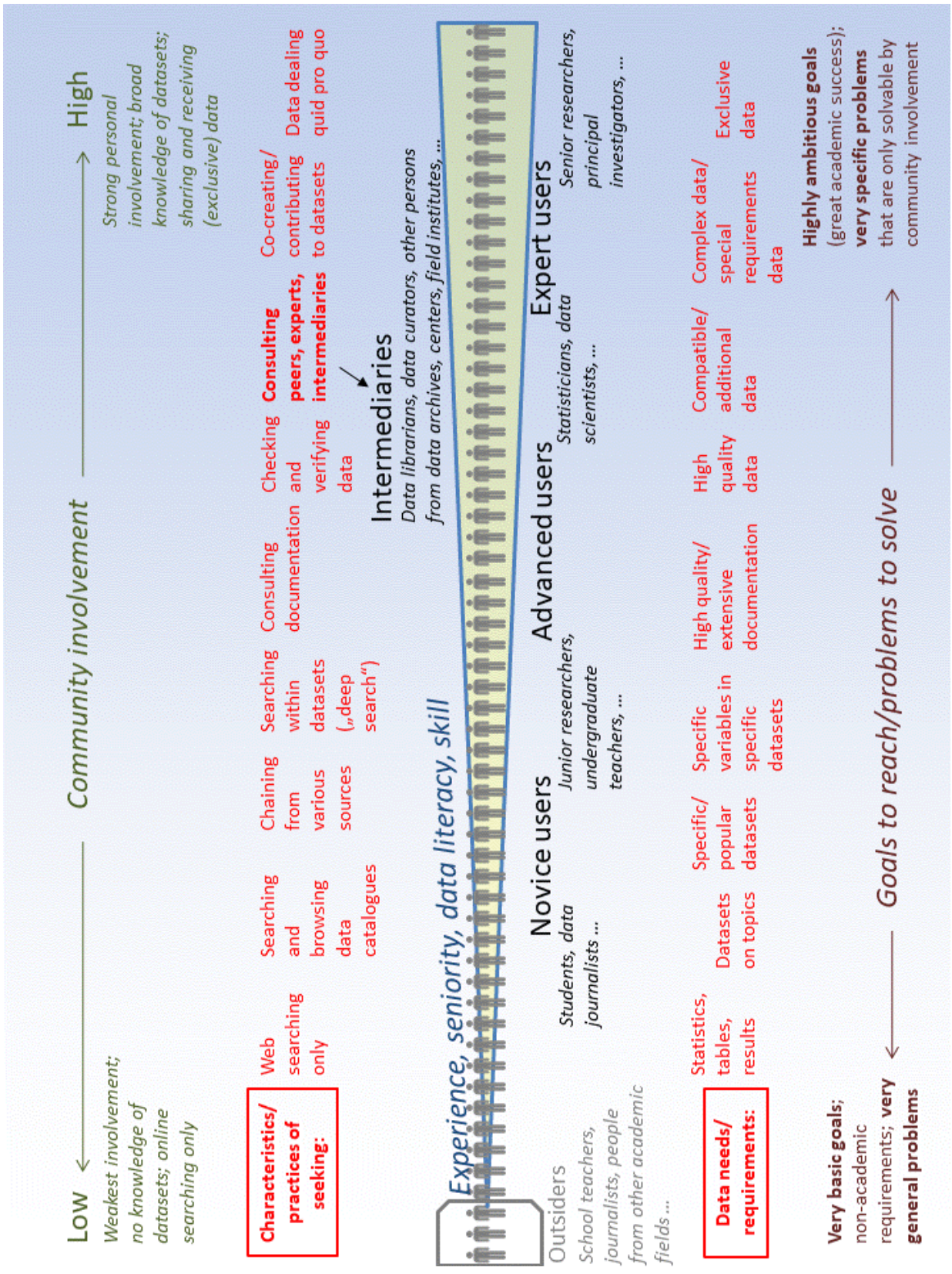
"Problem solving by community involvement" is a working phrase for a core conceptual category in the Grounded Theory of information seeking behaviour of secondary data users.

The category is connected to the previously introduced understanding of information seeking as goal-oriented problem solving (Wilson 2002). Goals and problems of secondary data users are very diverse, as the interviews have shown. Goals that shine through in users' requests to data service are: getting published; success or prestige; following trends; researching remarkable topics; doing original research; learning how to work with data; graduating. Goals are influenced by background and skills of data users. In turn, goals influence requirements, which have direct effect on seeking behaviours (e.g. in relevance judgement). Problems or barriers that people face when they try to reach their goals are: lack of recent data; lack of suitable data; lack of information on data; lack of skill; problems with data access; problems with infrastructure; legal barriers; problems with data quality.

Goals, requirements, and barriers trigger people to seek information that helps them to resolve the problematic situation (understood as discrepancy between their life-world and encountered phenomena, see Wilson). How they proceed to seek the information that they need is the core question of this investigation.

The interviews suggest that a significant factor in problem resolution is personal interaction with others. In particular, for certain datasets there seems to be a vital community that people can join to improve their own outcomes as well as outcomes of others who are working on similar problems or just with the same dataset (see memo "dataset communities"). Data-related communities are not necessarily dataset specific. Data communities may consist of peers (students, colleagues), supervisors/teachers, data professionals such as reference persons (intermediaries). A person can be part of various communities and in different roles. Being part of one or more communities in this sense increases the individual capability to solve problems when looking for data. This means that information seeking behaviour with regard to survey data is influenced by community involvement (as are goals and requirements that induce and determine information seeking behaviour).

Annex 16: Diagram "Model of problem-solving by community involvement"



Annex 17: Introduction for Respondent Validation (German)

In meiner Studie untersuche ich das Informationssuchverhalten von Personen, die Umfragedaten suchen. Ich orientiere mich dabei an verschiedenen etablierten Theorien der Informationsverhaltensforschung. Informationssuchverhalten kann man vor dem Hintergrund dieser Theorien als zielorientiertes Lösen von Problemen verstehen und betrachten. Das ist der Ausgangspunkt meiner Untersuchung. Der Fokus liegt dabei darauf herauszufinden, welche Einflussfaktoren diesen Prozess wesentlich beinflussen. Neben den individuellen Zielen und Problemen gibt es weitere Faktoren, denen ich in meinen Interviews auf den Grund gehen wollte.

Die Interviews haben mir sehr dabei geholfen, das Datensuchverhalten besser zu verstehen. Die wichtigste Erkenntnis, auf der auch meine Theorie beruht, ist dass es in der Umfrageforschung verschiedene Communities gibt, innerhalb derer nicht nur die Produktion und Nutzung von Daten, sondern auch die Suche nach Daten stattfindet. Meine Kernhypothese ist, dass die Einbindung in die Community einen wesentlichen Einfluss auf den Erfolg bei der Datensuche hat. Meine Theorie „problem-solving by community involvement“ besagt unter anderem, dass das Erreichen ambitionierter Ziele und das Lösen komplexer Probleme besser oder im äußersten Fall sogar ausschließlich gelingen, wenn man eine möglichst hohe Community-Einbindung hat.

Zurückgebunden an bisherige Forschung zum Thema Informationsverhalten bedeutet das, dass im Falle der Datensuche informelle Informationskanäle eine besonders wichtige Rolle spielen.

Das Diagramm beschreibt die verschiedenen Aspekte der Theorie des „problem-solving by community involvement wie folgt:

Personen, die nach Daten suchen können entlang eines Spektrums unterschiedlicher Erfahrung, Seniorität, Datenkompetenz und anderer Fähigkeiten beschrieben werden.

Ganz links finden sich Personen, die als Außenseiter beschrieben werden können, denn sie haben keine Erfahrung mit Umfragedaten, keine Datenkompetenz und keine Fähigkeiten oder Kenntnisse in der Nutzung und Analyse von Daten.

Ganz rechts befinden sich dagegen die absoluten Insider der Umfragedatennutzung. Sie sind etablierte Forscherinnen und Forscher mit viel Erfahrung und sehr guten Kenntnissen und Fähigkeiten. Zu diesen Personen gehören auch die Primärforscher in großen Survey-Programmen. Die Außenseiter sind nicht in die Community eingebunden, die Insider dagegen sehr stark.

Im Hinblick auf das Datensuchverhalten unterscheiden sich diese Personen sehr. Entlang des Spektrums haben sie verschiedene Bedürfnisse oder Ansprüche an Daten und wenden verschiedene Praktiken der Suche an.

Diese Unterschiede korrespondieren mit ihren verschiedenartigen Zielen und Problemen. Ziele und Probleme bestimmen die Informationsbedürfnisse und Ansprüche an die Daten.

Der Grad des Community involvements korreliert mit dem Suchverhalten bzw. den Suchpraktiken.

Nur ein großes Community involvement gewährt die Anwendbarkeit bestimmter Praktiken und damit auch die Lösung komplexer Probleme und das Erreichen ambitionierter Ziele.

Annex 18: Questionnaire (English)

ID	Q01
Filter:	Ask all
Instruction:	Single answer
Label:	Language

◌ English
◌ German

Introduction:

This survey was designed to gather knowledge about how people search and use survey data. The results of this survey will be used to improve data services for users. The survey is part of a PhD project by Tanja Friedrich (GESIS and Humboldt-Universität zu Berlin). You can learn more about this project in this consent form. The consent form also informs you about the handling and processing of the data that is collected with this survey. By clicking "START SURVEY" at the end of this page, you agree that your contribution is included in this research. It will take about 10 to 15 minutes to complete the survey. Thank you very much for your participation.

START SURVEY

ID	Q02
Filter:	Ask all
Instruction:	Single answer
Label:	Use of data

Q02 - We start with a few questions on your past and present use of survey data. Have you ever used survey data for your work or for your studies?

◌ Yes, I have used survey data.
◌ No, but I have used other research data. Please specify, which kind of research data you have previously used:
◌ No, I have never used any survey data or other research data.

ID	Q03
Filter:	Ask only those who have used survey data (“Yes, …” in Q2)
Instruction:	Single answer
Label:	Data analysis

Q03 - Have you ever performed statistical analyses using survey data?

◯ Yes, once or twice.
◯ Yes, more than twice.
◯ No, never.

ID	Q04
Filter:	Ask only those who have done statistical analyses ("Yes"= item 1 or 2 in Q3)
Instruction:	Multiple answers; randomize; anchor last item
Label:	Methodological skills

Q04 - What methods have you used for survey data analysis so far? I have used ...
Multiple answers are possible

☐ ... basic methods of analysis (such as counting, frequencies, distributions or other univariate analyses).
☐ ... advanced methods of analysis (such as cross tabulation or other bivariate analyses).
☐ ... expert methods of analysis (such as multiple regression or other multivariate analyses).
☐ ... other methods, please specify:

ID	Q05
Filter:	Ask only those who have used survey data or other data (item 1 or 2 in Q2)
Instruction:	Multiple answers; randomize; anchor last item
Label:	Software skills

Q05 - What software have you used to analyse data?
Multiple answers are possible

☐ Excel
☐ SPSS
☐ Stata
☐ SAS
☐ MPlus
☐ R
☐ I have used other software for data analysis. The software I have used is:

ID	Q06
Filter:	Ask only those who have used survey data ("Yes, …" in Q02)
Instruction:	Multiple answers; randomize; anchor last item
Label:	Goals/purpose

Q06 - For what purposes did you use survey data in the past two years?I have used ...
Multiple answers are possible

☐ ... survey data for my thesis (bachelor thesis, master thesis, PhD thesis, etc.).

- ... survey data to support a non-scientific publication (book, newspaper article, etc.).
- ... survey data for a scientific publication (journal article, conference publication, etc.).
- ... survey data to support a policy paper or strategy paper.
- ... survey data to practice or to learn how to work with survey data.
- ... survey data to look at it and come up with an interesting research question.
- ... existing measures (questions, scales etc.) for my own survey.
- ... a specific dataset to replicate results of a study with the same dataset.
- ... survey data for teaching.
- ... survey data for something else (please specify):

ID	Q07/08[31]
Filter:	Ask all
Instruction:	Multiple answers; randomize; Image as answer
Label:	Known data/closed

Q07/08 - Have you ever heard of the following surveys? Please select all the surveys that you have heard of by clicking on the survey logo.

Multiple answers are possible

- ALLBUS (Allgemeine Bevölkerungsumfrage der Sozialwissenschaften)
- BHPS (British Household Panel Survey)
- BSA (British Social Attitudes Survey)
- CILS4EU (Children of Immigrants' Longitudinal Survey)
- CSES (Comparative Study of Electoral Systems)
- EES (European Election Studies)
- ESS (European Social Survey)
- Eurobarometer
- Eurofound European Working Conditions Survey
- Eurofound European Quality of Life Survey
- EVS (European Values Study)
- GESIS Panel
- GIP (German Internet Panel)
- GLES (German Longitudinal Election Study)
- GMF (Gruppenbezogene Menschenfeindlichkeit)
- GSS (General Social Survey)
- ISSP (International Social Survey Programme)
- NEPS (Nationales Bildungspanel)
- Pairfam (Panel Analysis of Intimate Relationships and Family Dynamics)
- PIAAC (Programme for the International Assessment of Adult Competencies)
- PISA (Programme for International Student Assessment)
- SHARE (Survey of Health, Ageing and Retirement in Europe)
- Shell Jugendstudie
- SOEP (Sozio-Oekonomisches Panel)

[31] Double IDs are used for questions that were administered in two different layouts for PC and mobile interfaces. In the software, the questions for both groups were treated as separate questions, resulting in two separate question IDs.

☐ WVS (World Values Study)

ID	Q09
Filter:	Ask all
Instruction:	Single answer
Label:	Known data/open

Q09 - What other important surveys do you know?

○ Fill in all other surveys that you know, separate with comma (,):
○ I don't know any other surveys.

ID	Q10
Filter:	Ask only those who ticked at least one item in Q07/08 and/or added at least one survey programme in Q09
Instruction:	Multiple answers; randomize; anchor last two items
Label:	Sources of known data

Q10 - Where do you know these survey programmes from?
Multiple answers are possible

☐ Web searches (e.g. Google, yahoo, bing).
☐ Dataset search engines (e.g. Google Dataset Search, Elsevier DataSearch, DataCite Search).
☐ Online catalogues of data archives (like figshare, Zenodo, or research data centres).
☐ Journal articles.
☐ Teachers/professors or supervisors.
☐ Colleagues or friends.
☐ Library services.
☐ Social media contacts (e.g. Facebook, ResearchGate, LinkedIn).
☐ Talks at conferences.
☐ Textbooks and other books.
☐ The media (tv, radio, newspaper).
☐ I am a principal investigator in one or more of these programmes.
☐ I know these survey programmes from other sources (please specify):

ID	Q11
Filter:	Ask all
Instruction:	Single answer
Label:	Seeking data

Q11 - Now we have a few questions on how you usually search and find suitable survey data for your work or your studies. In the past two years, have you searched for survey data that you could use for your work or your studies?

◯ Yes.
◯ No.

ID	Q12/13
Filter:	Ask only those who ticked “Yes” in Q11
Instruction:	5-point Likert scale: not important at all … very important; randomize
Label:	Requirements/closed

Q12/13 - When searching for these data, how important were each of the following requirements? Please indicate importance on a scale from 1 (not important at all) to 5 (very important). **The data should...**

	1 = not important at all	2	3	4	5 = very important
... be easy to understand (e.g., results, tables, or simple statistics).	◯	◯	◯	◯	◯
... be available free of charge.	◯	◯	◯	◯	◯
... fit my research question.	◯	◯	◯	◯	◯
... come from a specific survey.	◯	◯	◯	◯	◯
... be as new as possible.	◯	◯	◯	◯	◯
... be well documented with sufficient descriptive information.	◯	◯	◯	◯	◯
... be of high quality.	◯	◯	◯	◯	◯
... come from a longitudinal survey, because I wanted to study change over time.	◯	◯	◯	◯	◯
... come from an international survey, because I wanted to make comparisons between countries.	◯	◯	◯	◯	◯
... be compatible with other data that I already had.	◯	◯	◯	◯	◯
... contain spatial information.	◯	◯	◯	◯	◯
... not have been analysed (a lot) before.	◯	◯	◯	◯	◯

ID	Q14
Filter:	Ask only those who ticked “Yes” in Q11
Instruction:	Text input; optional
Label:	Requirements/open

Q14 - I had other important requirements (please specify):

ID	Q15
Filter:	Ask only those who ticked “Yes” in Q11
Instruction:	Multiple answers; randomize; anchor last item
Label:	Seeking/sources

Q15 - Which of the following sources do you use to find suitable data?
Multiple answers are possible

- ☐ Web searches (e.g. Google, yahoo, bing).
- ☐ Dataset search engines (e.g. Google Dataset Search, ElsevierDataSearch, DataCite Search).
- ☐ Online catalogues of data archives (like researchdata centres, figshare, Zenodo).
- ☐ Websites of census bureaus or bureaus of statistics.
- ☐ Relevant journal articles or other publications that mention or cite datasets.
- ☐ My professor/teacher or supervisor.
- ☐ Colleagues or friends.
- ☐ Librarians, data librarians, or other data specialists.
- ☐ Message boards or social media (Facebook, ResearchGate, LinkedIn etc.).
- ☐ I directly search datasets from surveys that I have worked with in the past.
- ☐ I have other ways of finding data (please specify):

ID	Q16
Filter:	Ask only those who ticked “Yes” in Q11
Instruction:	Maximum of 5 answers; randomize; anchor last item
Label:	Problems

Q16 - What are the main problems that you have encountered when finding or accessing survey data? Please give a **maximum of 5 answers**.
A maximum of 5 answers are possible

- ☐ I didn’t know where to find data.
- ☐ I didn’t know how to open or read the dataset.
- ☐ I didn’t have the knowledge to understand the content of the dataset.
- ☐ I couldn’t find data on my topic of interest.
- ☐ I couldn't find data on my population of interest.
- ☐ The data I found were too old.
- ☐ The data I found were of poor quality.
- ☐ Description or information on the data was insufficient.
- ☐ Description or information on the data was incorrect.
- ☐ I was denied access to data for legal or other reasons.
- ☐ I had other problems (please specify):
- ☐ I have never had problems finding or accessing survey data.

ID	Q17/18
Filter:	Ask all, except those who ticked “I have never had problems …” in Q16
Instruction:	5-point Likert scale: not important at all … very important; randomize
Label:	Problem solving/closed

Q17/18 - How do you deal with problems of finding and accessing survey data? Please indicate how important the following strategies of problem solving are for you on a scale from 1 (not important at all) to 5 (very important).

	1 = not important at all	2	3	4	5 = very important
I try to solve the problem by reading documentation and other information material (like code books, field reports, etc.).	○	○	○	○	○
I try to find help in online message boards or social media (Facebook, ResearchGate, LinkedIn, etc.).	○	○	○	○	○
I participate in training or a workshop that deals with this problem.	○	○	○	○	○
I visit a conference or another event that deals with the survey data that I want to work with.	○	○	○	○	○
I ask my professors/teachers or supervisors for help.	○	○	○	○	○
I ask colleagues or friends for help.	○	○	○	○	○
I ask data librarians or other data specialists for help.	○	○	○	○	○
I ask the person who collected the data for help.	○	○	○	○	○
I conduct my own survey.	○	○	○	○	○
I adjust my research question to avoid the problem.	○	○	○	○	○

ID	Q19
Filter:	Ask all, except those who ticked “I have never had problems …” in Q16
Instruction:	Text input; optional
Label:	Problem solving/other

Q19 - I have another important strategy (please specify):

ID	Q20
Filter:	Ask only those who have used survey data ("Yes, …" in Q02)
Instruction:	Single answer
Label:	Data collection

Q20 - In the last part of this survey, we would like to know more about your own survey data projects. Have you ever conducted a survey and produced survey data (either on your own or together with other people)?

◯ Yes.
◯ No.

ID	Q21
Filter:	Ask only those who have collected survey data ("Yes" in Q20)
Instruction:	Single answer
Label:	Data sharing/if

Q21 - Have you ever shared data from your own survey (or from a survey that you have conducted with others)? Sharing data means that you have provided someone else with your dataset, either in person or through a website, a data archive, a library, a data repository or any other channel.

◯ Yes, I have shared survey data.
◯ No, I haven't shared survey data (yet).

ID	Q22
Filter:	Ask only those who have shared survey data ("Yes, …" in Q21)
Instruction:	Multiple answers; randomize; anchor last item
Label:	Data sharing/how

Q22 - How have you shared your survey datasets? Please think of any survey data that you have shared in the past. I have ...

Multiple answers are possible

☐ ... shared my data with a colleague or friend.
☐ ... shared my data upon request through social media (Facebook, ResearchGate, LinkedIn, etc.).
☐ ... published my data through the repository or website of my institution.
☐ ... published my data on my personal website.
☐ ... published my data on my social media page (Facebook, ResearchGate, LinekdIn, etc.).
☐ ... published my data on the website of the project that had produced the survey.
☐ ... acted as a principal investigator in a survey programme that produces data that are

generally made available for the research community.
☐ ... been required to provide my survey data to a journal or book publisher or to peer reviewers as part of the publishing process of a journal article or book (article).
☐ ... published survey data through an online repository or catalogue (like zenodo or figshare).
☐ ... published survey data through a data archive (like GESIS, UK Data Archive, ICPSR).
☐ ... shared survey data in another way (please specify):

ID	Q23/24
Filter:	Ask only those who ticked "Yes, …" in Q02
Instruction:	Multiple answers; randomize; anchor last item
Label:	Own contribution

Q23/24 - Some people who are working with survey data contribute to the creation, improvement, or dissemination of survey data for reuse in some way or another. Have you ever engaged in one or more of the following activities? **I have ...**

	No	Yes
... contributed one or more questions to an access panel.	◯	◯
... found an error in a dataset and reported it to the distributor of the data (e.g. to the data archive or the principle investigator) or shared a corrected version with the distributor.	◯	◯
... suggested an improvement of a dataset to the distributor of the data (e.g. to the data archive or the principle investigator) or shared an improved version of a dataset wit the distributor.	◯	◯
... shared with others a syntax file that I had created, so they could reuse it.	◯	◯
... shown or taught people how to find survey data or how to work with a specific survey dataset.	◯	◯
... helped people who had problems with a dataset or pointed them to persons who could help them.	◯	◯
... shared a publicly available dataset that I had not created by myself with other people.	◯	◯
... shared an access-restricted dataset with other people.	◯	◯
… been a consultant or member of an advisory board for a project or institution	◯	◯

	No	Yes
that creates and publishes survey data.		
... contributed to the creation, improvement, or dissemination of survey data in another way, (please specify):	◌	◌

ID	Q25
Filter:	Ask all
Instruction:	Number input
Label:	Age

Q25 - You are almost done with this survey. We just need some further information on you to be able to categorize your answers. How old are you?

◌ younger than 21 years
◌ 21 to 30 years
◌ 31 to 40 years
◌ 41 to 50 years
◌ 51 to 60 years
◌ 61 to 70 years
◌ older than 71 years
◌ No answer

ID	Q26
Filter:	Ask all
Instruction:	Single answer
Label:	Gender

Q26 - Please indicate your gender:

◌ Female
◌ Male
◌ Other
◌ No answer

ID	Q27
Filter:	Ask all
Instruction:	Single answer; drop down ISO 3166-1 countries
Label:	Country of residence

Q27 - What is your current country of residence?
Please choose from the list below.

◌ Afghanistan

- ◌ Aland Islands
- ◌ Albania
- ◌ Algeria
- ◌ American Samoa
- ◌ Andorra
- ◌ Angola
- ◌ Anguilla
- ◌ Antarctica
- ◌ Antigua and Barbuda
- ◌ Argentina
- ◌ Armenia
- ◌ Aruba
- ◌ Australia
- ◌ Austria
- ◌ Azerbaijan
- ◌ Bahamas (the)
- ◌ Bahrain
- ◌ Bangladesh
- ◌ Barbados
- ◌ Belarus
- ◌ Belgium
- ◌ Belize
- ◌ Benin
- ◌ Bermuda
- ◌ Bhutan
- ◌ Bolivia (Plurinational State of)
- ◌ Bonaire, Sint Eustatius and Saba
- ◌ Bosnia and Herzegovina
- ◌ Botswana
- ◌ Bouvet Island
- ◌ Brazil
- ◌ British Indian Ocean Territory (the)
- ◌ Brunei Darussalam
- ◌ Bulgaria
- ◌ Burkina Faso
- ◌ Burundi
- ◌ Cabo Verde
- ◌ Cambodia
- ◌ Cameroon
- ◌ Canada
- ◌ Cayman Islands (the)
- ◌ Central African Republic (the)
- ◌ Chad
- ◌ Chile
- ◌ China
- ◌ Christmas Island
- ◌ Cocos (Keeling) Islands (the)

- ◌ Colombia
- ◌ Comoros (the)
- ◌ Congo (the Democratic Republic of the)
- ◌ Congo (the)
- ◌ Cook Islands (the)
- ◌ Costa Rica
- ◌ Côte d'Ivoire
- ◌ Croatia
- ◌ Cuba
- ◌ Curaçao
- ◌ Cyprus
- ◌ Czechia
- ◌ Denmark
- ◌ Djibouti
- ◌ Dominica
- ◌ Dominican Republic (the)
- ◌ Ecuador
- ◌ Egypt
- ◌ El Salvador
- ◌ Equatorial Guinea
- ◌ Eritrea
- ◌ Estonia
- ◌ Eswatini
- ◌ Ethiopia
- ◌ Falkland Islands (the) [Malvinas]
- ◌ Faroe Islands (the)
- ◌ Fiji
- ◌ Finland
- ◌ France
- ◌ French Guiana
- ◌ French Polynesia
- ◌ French Southern Territories (the)
- ◌ Gabon
- ◌ Gambia (the)
- ◌ Georgia
- ◌ Germany
- ◌ Ghana
- ◌ Gibraltar
- ◌ Greece
- ◌ Greenland
- ◌ Grenada
- ◌ Guadeloupe
- ◌ Guam
- ◌ Guatemala
- ◌ Guernsey
- ◌ Guinea
- ◌ Guinea-Bissau

- ◌ Guyana
- ◌ Haiti
- ◌ Heard Island and McDonald Islands
- ◌ Holy See (the)
- ◌ Honduras
- ◌ Hong Kong
- ◌ Hungary
- ◌ Iceland
- ◌ India
- ◌ Indonesia
- ◌ Iran (Islamic Republic of)
- ◌ Iraq
- ◌ Ireland
- ◌ Isle of Man
- ◌ Israel
- ◌ Italy
- ◌ Jamaica
- ◌ Japan
- ◌ Jersey
- ◌ Jordan
- ◌ Kazakhstan
- ◌ Kenya
- ◌ Kiribati
- ◌ Korea (the Democratic People's Republic of)
- ◌ Korea (the Republic of)
- ◌ Kuwait
- ◌ Kyrgyzstan
- ◌ Lao People's Democratic Republic (the)
- ◌ Latvia
- ◌ Lebanon
- ◌ Lesotho
- ◌ Liberia
- ◌ Libya
- ◌ Liechtenstein
- ◌ Lithuania
- ◌ Luxembourg
- ◌ Macao
- ◌ Macedonia (the former Yugoslav Republic of)
- ◌ Madagascar
- ◌ Malawi
- ◌ Malaysia
- ◌ Maldives
- ◌ Mali
- ◌ Malta
- ◌ Marshall Islands (the)
- ◌ Martinique
- ◌ Mauritania

- ◌ Mauritius
- ◌ Mayotte
- ◌ Mexico
- ◌ Micronesia (Federated States of)
- ◌ Moldova (the Republic of)
- ◌ Monaco
- ◌ Mongolia
- ◌ Montenegro
- ◌ Montserrat
- ◌ Morocco
- ◌ Mozambique
- ◌ Myanmar
- ◌ Namibia
- ◌ Nauru
- ◌ Nepal
- ◌ Netherlands (the)
- ◌ New Caledonia
- ◌ New Zealand
- ◌ Nicaragua
- ◌ Niger (the)
- ◌ Nigeria
- ◌ Niue
- ◌ Norfolk Island
- ◌ Northern Mariana Islands (the)
- ◌ Norway
- ◌ Oman
- ◌ Pakistan
- ◌ Palau
- ◌ Palestine, State of
- ◌ Panama
- ◌ Papua New Guinea
- ◌ Paraguay
- ◌ Peru
- ◌ Philippines (the)
- ◌ Pitcairn
- ◌ Poland
- ◌ Portugal
- ◌ Puerto Rico
- ◌ Qatar
- ◌ Réunion
- ◌ Romania
- ◌ Russian Federation (the)
- ◌ Rwanda
- ◌ Saint Barthélemy
- ◌ Saint Helena, Ascension and Tristan da Cunha
- ◌ Saint Kitts and Nevis
- ◌ Saint Lucia

- ◌ Saint Martin (French part)
- ◌ Saint Pierre and Miquelon
- ◌ Saint Vincent and the Grenadines
- ◌ Samoa
- ◌ San Marino
- ◌ Sao Tome and Principe
- ◌ Saudi Arabia
- ◌ Senegal
- ◌ Serbia
- ◌ Seychelles
- ◌ Sierra Leone
- ◌ Singapore
- ◌ Sint Maarten (Dutch part)
- ◌ Slovakia
- ◌ Slovenia
- ◌ Solomon Islands
- ◌ Somalia
- ◌ South Africa
- ◌ South Georgia and the South Sandwich Islands
- ◌ South Sudan
- ◌ Spain
- ◌ Sri Lanka
- ◌ Sudan (the)
- ◌ Suriname
- ◌ Svalbard and Jan Mayen
- ◌ Sweden
- ◌ Switzerland
- ◌ Syrian Arab Republic
- ◌ Taiwan (Province of China)
- ◌ Tajikistan
- ◌ Tanzania, United Republic of
- ◌ Thailand
- ◌ Timor-Leste
- ◌ Togo
- ◌ Tokelau
- ◌ Tonga
- ◌ Trinidad and Tobago
- ◌ Tunisia
- ◌ Turkey
- ◌ Turkmenistan
- ◌ Turks and Caicos Islands (the)
- ◌ Tuvalu
- ◌ Uganda
- ◌ Ukraine
- ◌ United Arab Emirates (the)
- ◌ United Kingdom of Great Britain and Northern Ireland (the)
- ◌ United States Minor Outlying Islands (the)

◯ United States of America (the)
◯ Uruguay
◯ Uzbekistan
◯ Vanuatu
◯ Venezuela (Bolivarian Republic of)
◯ Viet Nam
◯ Virgin Islands (British)
◯ Virgin Islands (U.S.)
◯ Wallis and Futuna
◯ Western Sahara
◯ Yemen
◯ Zambia
◯ Zimbabwe

ID	Q28
Filter:	Ask all
Instruction:	Single answer
Label:	Degree

Q28 - What is your highest college or university degree?

◯ I have no college or university degree (yet).
◯ I have a bachelor degree (or equivalent).
◯ I have a master degree (or equivalent).
◯ I have a doctoral degree (PhD or equivalent).
◯ I have a postdoctoral degree (habilitation or equivalent).
◯ I have another degree. My highest degree is:

ID	Q29
Filter:	Ask all
Instruction:	Single answer
Label:	Job status

Q29 - What is your current job status? Only tick your main occupation. I am ...

◯ ... a fulltime student.
◯ ... employed (includes traineeships; temporary leave; etc.).
◯ ... self-employed or freelancer.
◯ ... unemployed.
◯ ... retired.
◯ No answer

ID	Q30
Filter:	Ask all who ticked "… a fulltime student" in Q29
Instruction:	Single answer
Label:	Stage of studies

Q30 - What stage of your studies are you currently in? I am ...

◯ ... a bachelor student (or equivalent).
◯ ... a master student (or equivalent).
◯ ... a PhD student (or equivalent).
◯ ... a postdoctoral researcher.
◯ ... in another stage of my studies (please specify):

ID	Q31
Filter:	Ask all who ticked "… employed", "… self-employed" or "No answer" in Q29
Instruction:	Single answer
Label:	Sector/branch/current

Q31 - Which economic sector or branch are you currently working in? I work ...

◯ ... in research, science, or technology (this includes: academic and non-academic research and development; academic/ university education).
◯ ... in information and communication (this includes: media, publishers, broadcasters, news agencies, telecommunication, IT services).
◯ ... in consultancy (this includes: public relations and communication, advertising, market research, legal services).
◯ ... in education (this includes all educational institutions, except academic/university education, see above).
◯ ... in public administration (this includes: government agency, economic and social policy of the community).
◯ ... in arts and culture (this includes: libraries, archives, museums).
◯ ... in a non-profit organisation (this includes: business and employers organisations, trade unions, religious organisations, political parties, politcal organisations, consumer organisations, youth organisations).
◯ ... in health care or social work (this includes: health care, residential care, social care, social work).
◯ ... in another branch. The branch I work in is:

ID	Q32
Filter:	Ask all who are employed or self-employed who work in research (item 1 in Q31) or have given no answer in Q31
Instruction:	Single answer
Label:	Current position

Q32 - What is your position with your current employer? Only tick your main occupation. I am ...

◌ ... a university or college professor (includes assistant professor, junior professor, or equivalent).
◌ ... a lecturer at a university or college (no professorship).
◌ ... a senior researcher or postdoc researcher.
◌ ... a junior researcher or PhD student.
◌ ... a research assistant (with bachelor degree or equivalent).
◌ ... a librarian.
◌ ... an administrator.
◌ ... in another position. My current position is:

ID	Q33
Filter:	Ask all who ticked “unemployed” or “retired” in Q29
Instruction:	Single answer
Label:	Sector/branch/last

Q33 - What is the economic sector or branch that you have last worked in? I have last worked ...

◌ ... in research, science, or technology (this includes: academic and non-academic research and development; academic/ university education).
◌ ... in information and communication (this includes: media, publishers, broadcasters, news agencies, telecommunication, IT services).
◌ ... in consultancy (this includes: public relations and communication, advertising, market research, legal services).
◌ ... in education (this includes all educational institutions, except academic/university education, see above).
◌ ... in public administration (this includes: government agency, economic and social policy of the community).
◌ ... in arts and culture (this includes: libraries, archives, museums).
◌ ... in a non-profit organisation (this includes: business and employers organisations, trade unions, religious organisations, political parties, politcal organisations, consumer organisations, youth organisations).
◌ ... in health care or social work (this includes: health care, residential care, social care, social work).
◌ ... in another branch. The branch I have last worked in is:

◯ I was a fulltime student.

ID	Q34
Filter:	Ask all who are retired or unemployed and have worked in research (item 1 in Q33) or have given no answer in Q33
Instruction:	Single answer
Label:	Last position

Q34 - What was your position with your last employer? Only tick your main occupation. I was ...

◯ ... a university or college professor (includes assistant professor, junior professor, or equivalent).
◯ ... a lecturer at a university or college (no professorship).
◯ ... a senior researcher or postdoc researcher.
◯ ... a junior researcher or PhD student.
◯ ... a research assistant (with bachelor degree or equivalent).
◯ ... a librarian.
◯ ... an administrator.
◯ ... in another position. My last position was:

ID	Q35
Filter:	Ask all except those who ticked "… a fulltime student" in Q29
Instruction:	Min 1, max 60
Label:	Professional experience

Q35 - How long have you been in your job or in similar jobs that you have had before?
Please enter the number of years in the box below.

[ ] years

☐ No answer.

ID	Q36
Filter:	Ask all who ticked item 1 in Q31 or Q33 and all students (item 1 in Q29)
Instruction:	Single answer; 3-level drop down
Label:	Discipline

Q36 - What is your main research discipline or field of study? Please choose the one that comes closest to what you are currently doing.

+ Humanities and Social Sciences
  - Ancient Cultures
  - History

- Fine Arts, Music, Theatre and Media Studies
- Library and Information Science
- Linguistics
- Social and Cultural Anthropology, Non-European Cultures, Jewish Studies and Religious Studies
- Theology
- Philosophy
- + Educational Research (please specify, if possible)
  - General Education and History of Education
  - General and Domain-Specific Teaching and Learning
  - Education Systems and Educational Institutions
  - Educational Research on Socialization, Welfare and Organisations
  - None of the above. My field is: _____
- Psychology
- + Social Sciences (please specify, if possible)
  - Sociological Theory
  - Empirical Social Research
  - Communication Sciences
  - Political Science
  - None of the above. My field is: _____
- + Economics (please specify, if possible)
  - Economic Theory
  - Economic Policy and Public Finance
  - Business Administration
  - Statistics and Econometrics
  - Economic and Social History
  - None of the above. My field is: _____
- Jurisprudence
- None of the above. My field is: _____

+ Life Sciences
  - Basic Research in Biology and Medicine
  - Plant Sciences
  - Zoology
  - Microbiology, Virology and Immunology
  - Medicine
  - Neurosciences
  - Agriculture, Forestry and Veterinary Medicine
  - None of the above. My field is: _____
+ Natural Sciences
  - Molecular Chemistry
  - Chemical Solid State and Surface Research
  - Physical and Theoretical Chemistry
  - Analytical Chemistry, Method Development (Chemistry)
  - Biological Chemistry and Food Chemistry
  - Polymer Research
  - Condensed Matter Physics
  - Optics, Quantum Optics and Physics of Atoms, Molecules and Plasmas

    - Particles, Nuclei and Fields
    - Statistical Physics, Soft Matter, Biological Physics, Nonlinear Dynamics
    - Astrophysics and Astronomy
    - Mathematics
    - Atmospheric Science, Oceanography and Climate Research
    - Geology and Palaeontology
    - Geophysics and Geodesy
    - Geochemistry, Mineralogy and Crystallography
    - Geography
    - Water Research
    - None of the above. My field is: _____
+ Engineering Sciences
    - Production Technology
    - Mechanics and Constructive Mechanical Engineering
    - Process Engineering, Technical Chemistry
    - Heat Energy Technology, Thermal Machines, Fluid Mechanics
    - Materials Engineering
    - Materials Science
    - Systems Engineering
    - Electrical Engineering and Information Technology
    - Computer Science
    - Construction Engineering and Architecture
    - None of the above. My field is: _____
- None of the above. My field is: _____

Annex 19: Questionnaire (German)

ID	Q01
Filter:	Ask all
Instruction:	Single answer
Label:	Language

◌ English
◌ German

Einleitung:

In dieser Befragung wird erforscht, wie Menschen nach Daten aus Bevölkerungsumfragen suchen und wie sie diese Umfragedaten verwenden. Die Ergebnisse dieser Befragung sollen dabei helfen, Datenservices für Nutzerinnen und Nutzer von Umfragedaten zu verbessern. Diese Befragung ist Teil des Promotionsprojekts von Tanja Friedrich, M.A. (GESIS und Humboldt-Universität zu Berlin). Weitere Informationen zu diesem Projekt erhalten Sie in der Einwilligungserklärung. Die Einwilligungserklärung informiert auch über die Verwendung und Verarbeitung der in dieser Befragung erhobenen Daten. Wenn sie die Befragung durch Klicken auf „UMFRAGE STARTEN“ beginnen, erklären Sie sich damit einverstanden, mit Ihren Antworten an dieser Studie teilzunehmen. Die Befragung dauert etwa 10 bis 15 Minuten. Vielen Dank für Ihre Teilnahme.

UMFRAGE STARTEN

ID	Q02
Filter:	Ask all
Instruction:	Single answer
Label:	Use of data

Q02 - Zu Beginn ein paar Fragen zu Ihrer Nutzung von Daten aus Bevölkerungsumfragen. Haben Sie jemals Daten aus Bevölkerungsumfragen oder Meinungsumfragen für Ihre Arbeit oder für Ihr Studium verwendet?

◌ Ja, ich habe bereits Umfragedaten verwendet.
◌ Nein, aber ich habe bereits andere Forschungsdaten verwendet. Bitte geben Sie an, welche Art von Forschungsdaten Sie bereits verwendet haben:
◌ Nein, ich habe nie Umfragedaten oder andere Forschungsdaten verwendet.

ID	Q03
Filter:	Ask only those who have used survey data (“Ja, …” in Q2)
Instruction:	Single answer
Label:	Data analysis

Q03 – Haben Sie bereits statistische Analysen mit Umfragedaten durchgeführt?

○ Ja, ein- oder zweimal.
○ Ja, häufiger als zweimal.
○ Nein, bisher nicht.

ID	Q04
Filter:	Ask only those who have done statistical analyses (“Ja”= item 1 or 2 in Q3)
Instruction:	Multiple answers; randomize; anchor last item
Label:	Methodological skills

Q04 – Welche Methoden haben Sie bisher für die Analyse von Umfragedaten verwendet? Ich habe ...
Mehrere Antworten sind möglich

☐ … grundlegende Analysemethoden angewandt (z.B. Häufigkeitsauszählungen, Häufigkeitsverteilungen oder andere univariate Analysen).
☐ … fortgeschrittene Analysemethoden angewandt (z.B. Kreuztabellen oder andere bivariate Analysen).
☐ … Expertenmethoden angewandt (z.B. multiple Regression oder andere multivariate Analysen).
☐ … andere Methoden angewandt. Bitte geben Sie an, welche Methoden der Datenanalyse Sie bereits angewendet haben:

ID	Q05
Filter:	Ask only those who have used survey data or other data (item 1 or 2 in Q2)
Instruction:	Multiple answers; randomize; anchor last item
Label:	Software skills

Q05 – Welche Computerprogramme haben Sie bereits für die Datenanalyse verwendet?
Mehrere Antworten sind möglich

☐ Excel
☐ SPSS
☐ Stata
☐ SAS
☐ MPlus
☐ R
☐ Andere Computerprogramme. Bitte geben Sie an, welche Computerprogramme:

ID	Q06
Filter:	Ask only those who have used survey data (“Ja, …” in Q02)
Instruction:	Multiple answers; randomize; anchor last item
Label:	Goals/purpose

Q06 - Wofür haben Sie in den letzten zwei Jahren Daten aus Bevölkerungs- und Meinungsumfragen verwendet? **Ich habe ...**
Mehrere Antworten sind möglich

☐ … Daten für meine Abschlussarbeit oder Qualifikationsarbeit (Bachelorarbeit, Masterarbeit, Dissertation, usw.) verwendet.
☐ … Daten als Grundlage für eine nicht-wissenschaftliche Veröffentlichung (z.B. für ein Buch, einen Zeitungsartikel usw.) verwendet.
☐ … Daten als Grundlage für eine wissenschaftliche Publikation (z.B. für einen Zeitschriften- oder Buchartikel, eine Konferenzpublikation, usw.).
☐ … Daten als Grundlage für ein Strategiepapier, Grundsatzpapier oder ähnliches verwendet.
☐ … Daten verwendet, um Analysemethoden zu üben oder zu erlernen.
☐ … mir Daten angeschaut, um aus ihnen eine interessante Forschungsfrage abzuleiten.
☐ … bereits existierende Messinstrumente (Fragen, Skalen usw.) für meine eigene Umfrage nachgenutzt.
☐ … einen bestimmten Datensatz verwendet, um Ergebnisse einer Studie zu replizieren.
☐ … Daten in der Lehre oder als Unterrichtsmaterial verwendet.
☐ … Datenfür einen anderen Zweck verwendet (bitte nennen Sie den Zweck):

ID	Q07/08[32]
Filter:	Ask all
Instruction:	Multiple answers; randomize; Image as answer
Label:	Known data/closed

Q07/Q08 – Haben Sie schon einmal von den folgenden Bevölkerungsumfragen gehört? Bitte wählen Sie alle Umfragen aus, von denen Sie schongehört haben. Klicken Sie dafür auf das jeweilige Logo.
Mehrere Antworten sind möglich

☐ ALLBUS (Allgemeine Bevölkerungsumfrage der Sozialwissenschaften)
☐ BHPS (British Household Panel Survey)
☐ BSA (British Social Attitudes Survey)
☐ CILS4EU (Children of Immigrants’ Longitudinal Survey)
☐ CSES (Comparative Study of Electoral Systems)
☐ EES (European Election Studies)

[32] Double IDs are used for questions that were administered in two different layouts for PC and mobile interfaces. In the software, the questions for both groups were treated as separate questions, resulting in two separate question IDs.

☐ ESS (European Social Survey)
☐ Eurobarometer
☐ Eurofound European Working Conditions Survey
☐ Eurofound European Quality of Life Survey
☐ EVS (European Values Study)
☐ GESIS Panel
☐ GIP (German Internet Panel)
☐ GLES (German Longitudinal Election Study)
☐ GMF (Gruppenbezogene Menschenfeindlichkeit)
☐ GSS (General Social Survey)
☐ ISSP (International Social Survey Programme)
☐ NEPS (Nationales Bildungspanel)
☐ Pairfam (Panel Analysis of Intimate Relationships and Family Dynamics)
☐ PIAAC (Programme for the International Assessment of Adult Competencies)
☐ PISA (Programme for International Student Assessment)
☐ SHARE (Survey of Health, Ageing and Retirement in Europe)
☐ Shell Jugendstudie
☐ SOEP (Sozio-Oekonomisches Panel)
☐ WVS (World Values Study)

ID	Q09
Filter:	Ask all
Instruction:	Single answer
Label:	Known data/open

Q09 – Welche anderen wichtigen Bevölkerungs- oder Meinungsumfragen fallen Ihnen ein?

○ Tragen Sie die Umfragen hier ein (mehrere Umfragen mit Komma trennen):
○ Ich kenne keine anderen Bevölkerungs- oder Meinungsumfragen.

ID	Q10
Filter:	Ask only those who ticked at least one item in Q07/08 and/or added at least one survey programme in Q09
Instruction:	Multiple answers; randomize; anchor last two items
Label:	Sources of known data

Q10 - Woher kennen Sie diese Umfragen? Ich kenne diese Umfragen ...
Mehrere Antworten sind möglich

☐ … durch Suchen im Internet (z.B. über Google, yahoo, bing).
☐ … durch Suchen in Datensuchmaschinen (z.B. Google Dataset Search, Elsevier DataSearch, DataCite Search).
☐ …durch Suchen in Online-Datenkatalogen (z.B. von figshare, Zenodo, Datenarchiven oder Forschungsdatenzentren).
☐ … aus Artikeln in Fachzeitschriften.

☐ … von meinen DozentInnen, ProfessorInnen oder Vorgesetzten.
☐ … KollegInnen oder Bekannten.
☐ … durch Bibliotheken.
☐ … über Kontakte in sozialen Medien (z.B. Facebook, ResearchGate, LinkedIn).
☐ …von Vorträgen auf Konferenzen.
☐ … aus Lehrbüchern oder anderen Büchern.
☐ … aus den Medien (Fernsehen, Radio, Tageszeitung).
☐ Ich bin Primärforscher oder Primärforscherin in einem oder mehreren dieser Umfrageprogramme.
☐ Ich kenne diese Umfragen aus anderen Quellen (bitte nennen Sie diese Quellen):

ID	**Q11**
Filter:	Ask all
Instruction:	Single answer
Label:	Seeking data

Q11 - Jetzt interessiert uns noch, wie Sie normalerweise nach Daten für Ihre Arbeit oder Ihr Studium suchen. Haben Sie in den letzten zwei Jahren nach Daten aus Bevölkerungs- oder Meinungsumfragen für Ihre Arbeit oder Ihr Studium gesucht?

○ Ja.
○ Nein.

ID	**Q12/13**
Filter:	Ask only those who ticked “Ja” in Q11
Instruction:	5-point Likert scale: überhaupt nicht wichtig … sehr wichtig; randomize
Label:	Requirements/closed

Q12/13 - Wie wichtig waren Ihnen bei der Suche nach diesen Daten die folgenden Anforderungen? Bitte geben Sie die Wichtigkeit auf einer Skala von 1 (überhaupt nicht wichtig) bis 5 (sehr wichtig) an. **Die Daten sollten …**

	1 = überhaupt nicht wichtig	2	3	4	5 = sehr wichtig
... leicht verständlich sein (z.B. Ergebnisse, Tabellen oder Statistiken).	○	○	○	○	○
... kostenfrei verfügbar sein.	○	○	○	○	○
… zu meiner Forschungsfrage passen.	○	○	○	○	○
… aus einer ganz bestimmten Studie stammen.	○	○	○	○	○
... so aktuell oder so neu wie möglich sein.	○	○	○	○	○
… gut und mit ausreichenden	○	○	○	○	○

Zusatzinformationen beschrieben oder dokumentiert sein.					
… von hoher Qualität sein.	◯	◯	◯	◯	◯
… aus einer Langzeitstudie stammen, da ich Veränderungen über die Zeit untersuchen wollte.	◯	◯	◯	◯	◯
… aus einer internationalen Studie stammen, da ich Vergleiche zwischen Ländern ziehen wollte.	◯	◯	◯	◯	◯
… anschlussfähig sein an andere Daten, die ich schon hatte.	◯	◯	◯	◯	◯
… geographische Informationen enthalten.	◯	◯	◯	◯	◯
… zuvor von niemand anderem oder kaum jemand anderem analysiert worden sein.	◯	◯	◯	◯	◯

ID	**Q14**
Filter:	Ask only those who ticked “Ja” in Q11
Instruction:	Text input; optional
Label:	Requirements/open

Q14 – Ich hatte andere wichtige Anforderungen (bitte nennen Sie Ihre Anforderungen):

ID	**Q15**
Filter:	Ask only those who ticked “Ja” in Q11
Instruction:	Multiple answers; randomize; anchor last item
Label:	Seeking/sources

Q15 - Wenn Sie nach Daten suchen, wo suchen sie dann?

Mehrere Antworten sind möglich

☐ Ich suche im Internet (z.B. mit Google, yahoo, bing).
☐ Ich suche in Datensuchmaschinen (z.B. Google Dataset Search, Elsevier DataSearch, DataCite Search).
☐ Ich suche in Online-Datenkatalogen (z.B. von Datenarchiven, Forschungsdatenzentren, figshare, Zenodo).
☐ Ich suche auf Webseiten von statistischen Ämtern.
☐ Ich suche in wissenschaftlichen Artikeln oder anderen Veröffentlichungen, in denen Datensätze genannt oder zitiert werden.
☐ Ich frage meine DozentInnen, ProfessorInnen oder Vorgesetzte nach passenden Daten.
☐ Ich frage KollegInnen oder Bekannte nach passenden Daten.
☐ Ich lasse mir von BibliothekarInnen, DatenbibliothekarInnen oder andere

DatenspezialistInnen helfen.
☐ Ich mache einen Eintrag in einem Online-Forum oder in einem Sozialen Netzwerk (Facebook, ResearchGate, LinkedIn usw.).
☐ Ich durchsuche direkt einen Datensatz aus einer Umfrage, mit der ich schongearbeitet habe.
☐ Ich gehe anders vor, wenn ich Daten suche (bitte beschreiben Sie, wie und wo Siesuchen):

ID	Q16
Filter:	Ask only those who ticked "Ja" in Q11
Instruction:	Maximum of 5 answers; randomize; anchor last item
Label:	Problems

Q16 - Was waren bisher Ihre Hauptprobleme, wenn Sie Daten suchen oder darauf zugreifen wollten? Bitte wählen Sie **maximal 5 Antworten** aus.
Maximal 5 Antworten sind möglich

☐ Ichwusste nicht, wo ich nach den Daten suchen sollte.
☐ Ichkonnte die Datei mit den Daten nicht öffnen oder nicht lesen.
☐ Mir fehlte das Wissen, um die Inhalte des Datensatzes zu verstehen.
☐ Ichkonnte keine Daten zu dem Thema finden, das mich interessierte.
☐ Ich konnte keine Daten zu der Personengruppe finden, die mich interessierte.
☐ Ich fand veraltete Daten, die nicht aktuell genug waren.
☐ Ich habe Daten von schlechter Qualität gefunden.
☐ Die Beschreibung oder die Informationen zu den Daten waren nicht ausreichend.
☐ Die Beschreibung oder die Informationen zu den Daten waren falsch.
☐ Ichdurfte aus rechtlichen oder aus anderen Gründen nicht auf Daten zugreifen.
☐ Ich hatte andere Schwierigkeiten (bitte beschreiben Sie, welche Schwierigkeiten Siehatten):
☐ Ich hatte bisher keine Schwierigkeiten, Daten zu finden oder darauf zuzugreifen.

ID	Q17/18
Filter:	Ask all, except those who ticked "Ich hatte bisher keine Schwierigkeiten …" in Q16
Instruction:	5-point Likert scale: überhaupt nicht wichtig … sehr wichtig; randomize
Label:	Problem solving/closed

Q17/18 - Wie gehen Sie mit solchen Schwierigkeiten um? Bitte geben Sie auf einer Skala von 1 (überhaupt nicht wichtig) bis 5 (sehr wichtig) an, wie wichtig die folgenden Strategien für Sie sind, wenn Sie Schwierigkeiten mit Daten oder mit der Datensuche haben.

	1 = überhaupt nicht wichtig	2	3	4	5 = sehr wichtig
Ich versuche, das Problem durch Lesen der Dokumentation oder anderer Informationsmaterialien (z.B. Codebuch, Methodenbericht) zu lösen.	○	○	○	○	○
Ich suche Hilfe in Online-Foren oder in Sozialen Medien (Facebook, ResearchGate, LinkedIn usw.).	○	○	○	○	○
Ich mache eine Fortbildung oder nehme an einem Workshop zu dem Thema teil.	○	○	○	○	○
Ich besuche eine Konferenz oder eine andere Veranstaltung zu der Umfrage, mit der ich arbeiten will.	○	○	○	○	○
Ich bitte Vorgesetzte oder meine DozentInnen/ProfessorInnen um Hilfe.	○	○	○	○	○
Ich bitte KollegInnen oder Bekannte um Hilfe.	○	○	○	○	○
Ich bitte DatenbibliothekarInnen oder andere SpezialistInnen um Hilfe.	○	○	○	○	○
Ich bitte die Person um Hilfe, die die Daten erhoben hat.	○	○	○	○	○
Ich führe selbst eine Umfrage durch.	○	○	○	○	○
Ich passe meine Forschungsfrage an die Situation an, um das Problem zu umgehen.	○	○	○	○	○

ID	**Q19**
Filter:	Ask all, except those who ticked “Ich hatte bisher keine Schwierigkeiten …” in Q16
Instruction:	Text input; optional
Label:	Problem solving/other

Q19 – Ich habe eine andere wichtige Strategie (bitte beschreiben Sie Ihre Strategie):

ID	**Q20**
Filter:	Ask only those who have used survey data (“Ja, …” in Q02)
Instruction:	Single answer
Label:	Data collection

Q20 - Im letzten Teil der Befragung wollen wir noch mehr über Ihre eigenen

Umfrageprojekte erfahren. Haben Sie schon einmal eine Umfrage durchgeführt und Umfragedaten erzeugt (entweder alleine oder zusammen mit anderen Personen)?

○ Ja.
○ Nein.

ID	Q21
Filter:	Ask only those who have collected survey data ("Ja" in Q20)
Instruction:	Single answer
Label:	Data sharing/if

Q21 – Haben Sie Daten aus ihrer eigenen Umfrage (oder aus einer Umfrage, die Sie zusammenmit anderen durchgeführt haben) schon einmal mit anderen geteilt? Damit ist gemeint, dass sie jemand anderem Ihren Datensatz zur Verfügung gestellt haben, entweder auf direktem Weg (persönlich) oder über eine Webseite, ein Datenarchiv, eine Bibliothek, einen Online-Datendienst oder auf einem anderen Weg.

○ Ja, ich habe meine Daten schon mit anderen geteilt.
○ Nein, ich habe meine Daten (bisher) nicht mit anderen geteilt.

ID	Q22
Filter:	Ask only those who have shared survey data ("Ja, …" in Q21)
Instruction:	Multiple answers; randomize; anchor last item
Label:	Data sharing/how

Q22 - Wie haben Sie Ihre Daten mit anderen geteilt? Bitte denken Sie dabei an alle Ihre Datensätze, die Sie in der Vergangenheit mit anderen geteilt haben. Ich habe ...
Mehrere Antworten sind möglich

☐ … meine Daten mit KollegInnen oder Bekannten geteilt.
☐ … meine Daten auf Anfrage über Soziale Medien geteilt (Facebook, ResearchGate, LinkedIn usw.).
☐ … meine Daten über die Webseite oder das Datenrepositorium der Einrichtung, in der ich arbeite, veröffentlicht.
☐ … meine Daten auf meiner privaten Webseite veröffentlicht.
☐ … meine Daten auf meiner Seite in einem Sozialen Netzwerk veröffentlicht (Facebook, ResearchGate, LinkedIn usw.).
☐ … meine Daten auf der Webseite des Projekts, in dem die Umfrage durchgeführt wurde, veröffentlicht.
☐ … als Primärforscher ein großes Umfrageprogramm mit durchgeführt, dessen Daten generell für die Forschungsgemeinschaft zur Verfügung gestellt werden.
☐ … im Rahmen der Veröffentlichung eines Zeitschriftenaufsatzes oder Buchartikels meine Daten einem Verlag oder den GutachterInnen mit übergeben müssen.
☐ … meine Daten über einen Online-Datendienst veröffentlicht (z.B. Zenodo oder figshare).
☐ … meine Daten zur Veröffentlichung an ein Datenarchiv übergeben (z.B. an GESIS, UK Data

Archive, ICPSR).
☐ … meine Daten auf andere Weise geteilt (bitte beschreiben Sie, wie Sie ihre Datengeteilt haben):

ID	Q23/24
Filter:	Ask only those who ticked “Ja, …” in Q02
Instruction:	Multiple answers; randomize; anchor last item
Label:	Own contribution

Q23/24 - Manche Personen, die mit Daten aus Bevölkerungs- oder Meinungsumfragen arbeiten, tragen in der einen oder anderen Weise zur Erstellung, Verbesserung oder Verbreitung von nachnutzbaren Umfragedaten bei. Wie ist das bei Ihnen? **Ich habe ...**

	Nein	Ja
… eine oder mehrere Fragen zum Fragebogen einer Panel-Umfrage beigetragen.	○	○
… einen Fehler in einem Datensatz gefunden und den Vertreiber des Datensatzes (z.B. Datenarchiv, PrimärforscherIn) darüber informiert oder ihm eine korrigierte Version zur Verfügung gestellt.	○	○
… dem Vertreiber eines Datensatzes (z.B. Datenarchiv, PrimärforscherIn) einen Vorschlag für die Verbesserung eines Datensatzes gemacht oder ihm eine verbesserte Version des Datensatzes zur Verfügung gestellt.	○	○
… einen Programmcode (oder Programmiersyntax) für einen Datensatz erstellt und diesen Code zur Nachnutzung an andere Personen weitergegeben.	○	○
… anderen Personen gezeigt oder sie darin unterrichtet, wie man Daten findet oder wie man mit einem bestimmten Datensatz arbeitet.	○	○
… anderen Personen geholfen, wenn sie Probleme mit einem Datensatz hatten (oder ihnen gesagt, wer ihnen mit diesem Problem helfen kann).	○	○
… einen frei verfügbaren Datensatz, den ich nicht selbst erstellt hatte, an andere Personen weitergegeben.	○	○
… einen nicht frei zugänglichen	○	○

	Nein	Ja
(zugangsbeschränkten) Datensatz, den ich nicht selbst erstellt hatte, an andere Personen weitergegeben.		
… als GutachterIn oder Beirat/Beirätin fungiert für ein Projekt oder ein Institut, das Umfragedaten erhebt und veröffentlicht.	○	○
… in anderer Weise zur Erstellung, Verbesserung oder Verbreitung von nachnutzbaren Umfragedaten beigetragen. Bitte erläutern Sie, wie Sie beigetragen haben:	○	○

ID	Q25
Filter:	Ask all
Instruction:	Number input
Label:	Age

Q25 - Sie sind fast am Ende der Befragung angekommen. Wir brauchen nur noch ein paar Informationen von Ihnen, die uns dabei helfen, Ihre Antworten einzuordnen. Wie alt sind Sie?

○ jünger als 21 Jahre
○ zwischen 21 und 30 Jahren
○ zwischen 31 und 40 Jahren
○ zwischen 41 und 50 Jahren
○ zwischen 51 und 60 Jahren
○ zwischen 61 und 70 Jahren
○ älter als 71 Jahre
○ Keine Antwort

ID	Q26
Filter:	Ask all
Instruction:	Single answer
Label:	Gender

Q26 - Bitte nennen Sie uns Ihr Geschlecht

○ Weiblich
○ Männlich
○ Divers
○ Keine Antwort

ID	Q27
Filter:	Ask all
Instruction:	Single answer; drop down ISO 3166-1 countries
Label:	Country of residence

Q27 - In welchem Land leben Sie derzeit?

Bitte wählen sie Ihr Land in der Liste aus.

- ○ Afghanistan
- ○ Ägypten
- ○ Aland
- ○ Albanien
- ○ Algerien
- ○ Amerikanische Jungferninseln
- ○ Amerikanisch-Samoa
- ○ Andorra
- ○ Angola
- ○ Anguilla
- ○ Antarktika
- ○ Antigua und Barbuda
- ○ Äquatorialguinea
- ○ Argentinien
- ○ Armenien
- ○ Aruba
- ○ Aserbaidschan
- ○ Äthiopien
- ○ Australien
- ○ Bahamas
- ○ Bahrain
- ○ Bangladesch
- ○ Barbados
- ○ Belarus (Weißrussland)
- ○ Belgien
- ○ Belize
- ○ Benin
- ○ Bermuda
- ○ Bhutan
- ○ Bolivien
- ○ Bonaire, Sint Eustatius und Saba (Niederlande)
- ○ Bosien und Herzegowina
- ○ Botswana
- ○ Bouvetinsel
- ○ Brasilien
- ○ Britische Jungferninseln
- ○ Britisches Territorium im indischen Ozean
- ○ Brunei Darussalam

- ○ Bulgarien
- ○ Burkina Faso
- ○ Burundi
- ○ Chile
- ○ China, Volksrepublik
- ○ Cookinseln
- ○ Costa Rica
- ○ Côte d'Ivoire (Elfentbeinküste)
- ○ Curaçao
- ○ Dänemark
- ○ Deutschland
- ○ Dominica
- ○ Dominikanische Republik
- ○ Dschibuti
- ○ Ecuador
- ○ El Salvador
- ○ Eritrea
- ○ Estland
- ○ Falklandinseln
- ○ Färöer
- ○ Fidschi
- ○ Finnland
- ○ Frankreich
- ○ Französische Süd- und Antarktisgebiete
- ○ Französisch-Guayana
- ○ Französisch-Polynesien
- ○ Gabun
- ○ Gambia
- ○ Georgien
- ○ Ghana
- ○ Gibraltar
- ○ Grenada
- ○ Griechenland
- ○ Grönland
- ○ Guadeloupe
- ○ Guam
- ○ Guatemala
- ○ Guernsey (Kanalinsel)
- ○ Guinea
- ○ Guinea-Bissau
- ○ Guyana
- ○ Haiti
- ○ Heard und McDonaldinseln
- ○ Honduras
- ○ Hongkong
- ○ Indien
- ○ Indonesien

- ○ Insel Man
- ○ Irak
- ○ Iran, Islamische Republik
- ○ Irland
- ○ Island
- ○ Israel
- ○ Italien
- ○ Jamaika
- ○ Japan
- ○ Jemen
- ○ Jersey (Kanalinsel)
- ○ Jordanien
- ○ Kaimaninseln
- ○ Kambodscha
- ○ Kamerun
- ○ Kanada
- ○ Kap Verde
- ○ Kasachstan
- ○ Katar
- ○ Kenia
- ○ Kirgisistan
- ○ Kiribati
- ○ Kokosinseln
- ○ Kolumbien
- ○ Komoren
- ○ Kongo, Demokratische Republik (ehem. Zaire)
- ○ Kongo, Republik (ehem. K.-Brazzaville)
- ○ Korea, Demokratische Volksrepublik (Nordkorea)
- ○ Korea, Republik (Südkorea)
- ○ Kroatien
- ○ Kuba
- ○ Kuwait
- ○ Laos, Demokratische Volksrepublik
- ○ Lesotho
- ○ Lettland
- ○ Libanon
- ○ Liberia
- ○ Libyen
- ○ Liechtenstein
- ○ Litauen
- ○ Luxemburg
- ○ Macau
- ○ Madagaskar
- ○ Malawi
- ○ Malaysia
- ○ Malediven
- ○ Mali

- ○ Malta
- ○ Marokko
- ○ Marschallinseln
- ○ Martinique
- ○ Mauretanien
- ○ Mauritius
- ○ Mayotte
- ○ Mazedonien
- ○ Mexiko
- ○ Mikronesien
- ○ Moldawien (Republik Moldau)
- ○ Monaco
- ○ Mongolei
- ○ Montenegro
- ○ Monserrat
- ○ Mosambik
- ○ Myanmar (Burma)
- ○ Namibia
- ○ Nauru
- ○ Nepal
- ○ Neukaledonien
- ○ Neuseeland
- ○ Nicaragua
- ○ Niederlande
- ○ Niger
- ○ Nigeria
- ○ Niue
- ○ Nördliche Marianen
- ○ Norfolkinsel
- ○ Norwegen
- ○ Oman
- ○ Österreich
- ○ Osttimor (Timor-Leste)
- ○ Pakistan
- ○ Palau
- ○ Panama
- ○ Papua-Neuguinea
- ○ Paraguay
- ○ Peru
- ○ Philippinen
- ○ Pitcairninseln
- ○ Polen
- ○ Portugal
- ○ Puerto Rico
- ○ Republik China (Taiwan)
- ○ Réunion
- ○ Ruanda

- ○ Rumänien
- ○ Russische Föderation
- ○ Saint-Barthélemy
- ○ Saint-Martin (franz. Teil)
- ○ Saint-Pierre und Miquelon
- ○ Salomonen
- ○ Sambia
- ○ Samoa
- ○ San Marino
- ○ São Tomé und Príncipe
- ○ Saudi-Arabien
- ○ Schweden
- ○ Schweiz (Confoederatio Helvetica)
- ○ Senegal
- ○ Serbien
- ○ Seychellen
- ○ Sierra Leaone
- ○ Simbabwe
- ○ Singapur
- ○ Sint Maarten (niederl. Teil)
- ○ Slowakei
- ○ Slowenien
- ○ Somalia
- ○ Spanien
- ○ Sri Lanka
- ○ St. Helena
- ○ St. Kitts und Nevis
- ○ St. Lucia
- ○ St. Vincent und die Grenadinen
- ○ Staat Palästina
- ○ Südafrika
- ○ Sudan
- ○ Südgeorgien und die Südlichen Sandwichinseln
- ○ Südsudan
- ○ Suriname
- ○ Svalbard und Jan Mayen
- ○ Swasiland
- ○ Syrien, Arabische Republik
- ○ Tadschikistan
- ○ Tansania, Vereinigte Republik
- ○ Thailand
- ○ Togo
- ○ Tokelau
- ○ Tonga
- ○ Trinidad und Tobago
- ○ Tschad
- ○ Tschechien

○ Tunesien
○ Türkei
○ Turkmenistan
○ Turks- und Caicoinseln
○ Tuvalu
○ Uganda
○ Ukraine
○ Ungarn
○ United States Minor Outlying Islands
○ Uruguay
○ Usbekistan
○ Vanuatu
○ Vatikanstadt
○ Venezuela
○ Vereinigte Arabische Emirate
○ Vereinigte Staaten von Amerika
○ Vereinigtes Königreich Großbritannien und Nordirland
○ Vietnam
○ Wallis und Futuna
○ Weihnachtsinsel
○ Westsahara
○ Zentralafrikanische Republik
○ Zypern

ID	Q28
Filter:	Ask all
Instruction:	Single answer
Label:	Degree

Q28 – Was ist Ihr höchster Hochschulabschluss?

○ Ich habe (noch) keinen Hochschulabschluss.
○ Ich habe einen Bachelor-Abschluss (oder vergleichbar).
○ Ich habe einen Master-Abschluss (oder vergleichbar, z.B. Magister-Abschluss, Diplom oder Staatsexamen).
○ Ich habe einen Doktorgrad.
○ Ich habe eine Habilitation abgeschlossen (oder andere postdoktorale Qualifikation, z.B. positive Zwischenevaluation der Junior-Professur).
○ Ich habe einen anderen Hochschulabschluss. Mein höchster Abschluss ist:

ID	Q29
Filter:	Ask all
Instruction:	Single answer
Label:	Job status

Q29 – Was ist Ihre derzeitige berufliche Stellung? Wählen Sie nur die Stellung aus, die hauptsächlich auf Sie zutrifft. Ich bin ...

○ ... Studentin oder Student in Vollzeit.
○ ... erwerbstätig (nicht selbständig; einschließlich Trainee, Volontariat, Referendariat etc.).
○ … selbständig erwerbstätig.
○ … nicht erwerbstätig oder arbeitslos.
○ … im Ruhestand.
○ Keine Antwort.

ID	Q30
Filter:	Ask all who ticked "… Studentin oder Student in Vollzeit" in Q29
Instruction:	Single answer
Label:	Stage of studies

Q30 - In welcher Phase Ihres Studiums bzw. Ihrer schulischen Ausbildung befinden Sie sich? Ich bin …

○ … im Bachelorstudium (oder vergleichbar).
○ … im Masterstudium (oder vergleichbar).
○ … DoktorandIn (oder vergleichbar).
○ … PostdoktorandIn.
○ … in einer anderen Phase (bitte benennen Sie die derzeitige Phase Ihres Studiums bzw. Ihrer schulischen Ausbildung):

ID	Q31
Filter:	Ask all who ticked "… erwerbstätig", "… selbständig erwerbstätig" or "Keine Antwort" in Q29
Instruction:	Single answer
Label:	Sector/branch/current

Q31 – In welchem Wirtschaftsbereich (Branche) arbeiten Sie derzeit? Ich arbeite ...

○ … im Bereich Forschung, Wissenschaft, Technologie (dazu gehören: universitäre und nicht-universitäre Forschung und Entwicklung; universitäre bzw. Hochschulbildung).
○ … im Bereich Information und Kommunikation (dazu gehören: Medien; Verlage; Rundfunk; Nachrichtenagenturen; Telekommunikation; IT-Dienstleistung).
○ … im Bereich Beratung (dazu gehören: PR- und Kommunikationsberatung; Werbung; Marktforschung; Rechtsberatung).
○ … im Bereich Erziehung (dazu gehören alle Erziehungs- und Bildungseinrichtungen; ausgenommen Hochschulbildung, s.o.).
○ … in der Öffentlichen Verwaltung (dazu gehören: Einrichtungen der öffentlichen Verwaltung, der öffentlichen Sicherheit der Rechtspflege und der Sozialversicherung).
○ … im Bereich Kunst und Kultur (dazu gehören: Bibliotheken; Archive; Museen).
○ … in einer Interessenvertretung oder Non-Profit-Organisation (dazu gehören:

Gewerkschaften; Arbeitgeberverbände; Industrieverbände; religiöse Vereinigungen; politische Parteien und Vereinigungen; Verbraucherorganisationen; kulturelle Organisationen; Jugendorganisationen).
○ … im Bereich Gesundheit und Soziale Arbeit (dazu gehören: Einrichtungen des Gesundheitswesens; Pflegeeinrichtungen; Einrichtungen der Sozialfürsorge).
○ … in einem anderen Wirtschaftsbereich. Der Wirtschaftsbereich in dem ich arbeite ist:

ID	Q32
Filter:	Ask all who are employed or self-employed who work in research (item 1 in Q31) or have given no answer in Q31
Instruction:	Single answer
Label:	Current position

Q32 – In welcher Position arbeiten Sie bei Ihrem derzeitigen Arbeitgeber? Bitte wählenSie nur Ihre Hauptbeschäftigung aus. Ich bin ...

○ … ProfessorIn an einer Hochschule (auch Juniorprofessur, Assistenzprofessur oder ähnliches).
○ … DozentInan einer Hochschule (keine Professur).
○ … PostdoktorandIn oder WissenschaftlerIn mit Leitungsfunktion (oder vergleichbar).
○ … NachwuchswissenschaftlerIn oder DoktorandIn (oder vergleichbar).
○ … Wissenschaftliche Hilfskraft (oder vergleichbar, z.B. mit Bachelorabschluss).
○ … BibliothekarIn.
○ … VerwaltungsmitarbeiterIn.
○ … in einer anderen Position. Meine derzeitige Position ist:

ID	Q33
Filter:	Ask all who ticked “nicht erwerbstätig oder arbeitslos” or “im Ruhestand” in Q29
Instruction:	Single answer
Label:	Sector/branch/last

Q33 - In welchem Wirtschaftsbereich (Branche) haben Sie zuletzt gearbeitet? Zuletzt arbeitete ich …

○ … im Bereich Forschung, Wissenschaft, Technologie (dazu gehören: universitäre und nicht-universitäre Forschung und Entwicklung; universitäre bzw. Hochschulbildung).
○ … im Bereich Information und Kommunikation (dazu gehören: Medien; Verlage; Rundfunk; Nachrichtenagenturen; Telekommunikation; IT-Dienstleistung).
○ … im Bereich Beratung (dazu gehören: PR- und Kommunikationsberatung; Werbung; Marktforschung; Rechtsberatung).
○ … im Bereich Erziehung (dazu gehören alle Erziehungs- und Bildungseinrichtungen; ausgenommen Hochschulbildung, s.o.).
○ … in der Öffentlichen Verwaltung (dazu gehören: Einrichtungen der öffentlichen Verwaltung, der öffentlichen Sicherheit der Rechtspflege und der Sozialversicherung).

○ … im Bereich Kunst und Kultur (dazu gehören: Bibliotheken; Archive; Museen).
○ … in einer Interessenvertretung oder Non-Profit-Organisation (dazu gehören: Gewerkschaften; Arbeitgeberverbände; Industrieverbände; religiöse Vereinigungen; politische Parteien und Vereinigungen; Verbraucherorganisationen; kulturelle Organisationen; Jugendorganisationen).
○ … im Bereich Gesundheit und Soziale Arbeit (dazu gehören: Einrichtungen des Gesundheitswesens; Pflegeeinrichtungen; Einrichtungen der Sozialfürsorge).
○ … in einem anderen Wirtschaftsbereich. Der Wirtschaftsbereich in dem ich zuletzt gearbeitet habe ist:
○ Ich war zuletzt StudentIn oder SchülerIn in Vollzeit.

ID	Q34
Filter:	Ask all who are retired or unemployed and have worked in research (item 1 in Q33) or have given no answer in Q33
Instruction:	Single answer
Label:	Last position

Q34 – In welcher Position waren Sie bei Ihrem letzten Arbeitgeber? Bitte wählen Sie nur Ihre Hauptbeschäftigung aus. Ich war ...

○ … ProfessorIn an einer Hochschule (auch Juniorprofessur, Assistenzprofessur oderähnliches).
○ … DozentIn an einer Hochschule (keine Professur).
○ … PostdoktorandIn oder WissenschaftlerIn mit Leitungsfunktion (oder vergleichbar).
○ … NachwuchswissenschaftlerIn oder DoktorandIn (oder vergleichbar).
○ … Wissenschaftliche Hilfskraft (oder vergleichbar, z.B. mit Bachelorabschluss).
○ … BibliothekarIn.
○ … VerwaltungsmitarbeiterIn.
○ … in einer anderen Position. Meine letzte Position war:

ID	Q35
Filter:	Ask all except those who ticked "… Studentin oder Student in Vollzeit" in Q29
Instruction:	Min 1, max 60
Label:	Professional experience

Q35 - Wie lange haben Sie bisher in diesem Beruf oder in ähnlichen Berufen gearbeitet?
Bitte geben Sie die Anzahl der Jahre an.

☐ Jahre

☐ Keine Antwort.

ID	Q36
Filter:	Ask all who ticked item 1 in Q31 or Q33 and all students (item 1 in Q29)
Instruction:	Single answer; 3-level drop down
Label:	Discipline

Q36 – Bitte wählen Sie hier Ihr Hauptforschungs- oder Studiengebiet aus. Wählen Sie das Fach aus, das Ihrer derzeitigen Tätigkeit am nächsten kommt.

- + Geistes-und Sozialwissenschaften
  - o Alte Kulturen
  - o Geschichte
  - o Kunst-, Musik-, Theater- und Medienwissenschaften
  - o Bibliotheks- und Informationswissenschaft
  - o Sprachwissenschaften
  - o Sozial- und Kulturanthropologie, Außereuropäische Kulturen, Judaistik und Religionswissenschaft
  - o Theologie
  - o Philosophie
  - + Erziehungswissenschaft und Bildungsforschung
    - o Allgemeine und Historische Pädagogik
    - o Allgemeines und fachbezogenes Lehren und Lernen
    - o Bildungssysteme und Bildungsinstitutionen
    - o Pädagogische Sozial- und Organisationsforschung
    - o Keines dieser Gebiete. Mein Hauptforschungs- oder Studiengebiet ist:
  - o Psychologie
  - + Sozialwissenschaften
    - o Soziologische Theorie
    - o Empirische Sozialforschung
    - o Publizistik und Kommunikationswissenschaft
    - o Politikwissenschaft
    - o Keines dieser Gebiete. Mein Hauptforschungs- oder Studiengebiet ist:
  - + Wirtschaftswissenschaften
    - o Wirtschaftstheorie
    - o Wirtschaftspolitik und Finanzwissenschaften
    - o Betriebswirtschaftslehre
    - o Statistik und Ökonometrie
    - o Wirtschafts- und Sozialgeschichte
    - o Keines dieser Gebiete. Mein Hauptforschungs- oder Studiengebiet ist:
  - o Rechtswissenschaften
  - o Keines dieser Gebiete. Mein Hauptforschungs- oder Studiengebiet ist:
- + Lebenswissenschaften
  - o Grundlagen der Biologie und Medizin
  - o Pflanzenwissenschaften
  - o Zoologie
  - o Microbiologie, Virologie und Immunologie
  - o Medizin
  - o Neurowissenschaft

- Agrar-, Forstwissenschaften und Tiermedizin
- Keines dieser Gebiete. Mein Hauptforschungs- oder Studiengebiet ist:

+ Naturwissenschaften
  - Molekülchemie
  - Chemische Festkörper- und Oberflächenforschung
  - Physikalische und Theoretische Chemie
  - Analytik, Methodenentwicklung (Chemie)
  - Biologische Chemie und Lebensmittelchemie
  - Polymerforschung
  - Physik der kondensierten Materie
  - Optik, Quantenoptik und Physik der Atome, Moleküle und Plasmen
  - Teilchen, Kerne und Felder
  - Statistische Physik, Weiche Materie, Biologische Physik, Nichtlineare Dynamik
  - Astrophysik und Astronomie
  - Mathematik
  - Atmosphären-, Meeres- und Klimaforschung
  - Geologie und Paläontologie
  - Geophysik und Geodäsie
  - Geochemie, Mineralogie und Kristallographie
  - Geographie
  - Wasserforschung
  - Keines dieser Gebiete. Mein Hauptforschungs- oder Studiengebiet ist:
+ Ingenieurwissenschaften
  - Produktionstechnik
  - Mechanik und Konstruktiver Maschinenbau
  - Verfahrenstechnik, Technische Chemie
  - Wärmeenergietechnik, Thermische Maschinen, Strömungsmechanik
  - Werkstofftechnik
  - Materialwissenschaft
  - Systemtechnik
  - Elektrotechnik und Informationstechnik
  - Informatik
  - Bauwesen und Architektur
  - Keines dieser Gebiete. Mein Hauptforschungs- oder Studiengebiet ist:
- Keines dieser Gebiete. Mein Hauptforschungs- oder Studiengebiet ist:

Annex 20: Consent Form (Web Survey)

Survey on data search and data use
Information and consent form

Principal Investigator:
Tanja Friedrich, M.A.
GESIS – Leibniz Institute for the Social Sciences
Unter Sachsenhausen 6-8
50667 Köln
Germany
E-Mail: tanja.friedrich@gesis.org

About this research project

This project investigates how users of survey data search and find reusable data. This research is necessary, because knowledge on how researchers, journalists and other people search, find and access data is important for the design of data libraries, data archives, data search engines, and other research data infrastructure. The present project aims at improving findability of research data and access to these data by making user-oriented recommendations for the design of research data infrastructures. In particular, the results of this study will be used to improve data services at the GESIS Leibniz Institute for the Social Sciences.

About the principal investigator

This research is part of a PhD project by Tanja Friedrich, M.A. The PhD supervisor is Prof. Vivien Petras, PhD, professor for information retrieval at the Berlin School of Library and Information Science, Humboldt-Universität zu Berlin. The PhD candidate Tanja Friedrich is also working as a researcher at the GESIS Leibniz Institute for the Social Sciences, a research institution that is funded by the German government. If you have questions, concerns or complaints about this study, please contact the principal investigator.

Data handling and data processing

The survey data will be collected with an online survey system. The online survey data sent over the Internet will be encrypted, no cookies will be used, and your IP addresses will not be collected. The information you provide will be stored on a password and firewall protected computer. Your responses will be recorded anonymously, that is to say they will not be collected or brought together with the contact information (name, e-mail) that we have used to contact you.

The results of the survey will be used for scientific publications and presentations. In all publications and presentations, results will be presented in a generalized and aggregated

way that assures that no personal identification of respondents will be possible from these reports.
The questionnaire contains optional questions on demographic information. Although it is highly improbable, this information could potentially be used to identify respondents. Therefore, this information will be anonymized (e.g. by grouping participants within broader groups according to their answers) before it will be archived. The anonymized data will be archived and made available for future research through the GESIS Data Archive for the Social Sciences. All prior data handling and data processing will be done exclusively by the principal investigator and GESIS data archive staff.

Your consent to participate
Your participation in this study is strictly voluntary. You may decline to participate, skip questions or withdraw your consent at any time. There is no risk or disadvantage in declining to participate or withdrawing your consent.
By clicking "start survey" on the first page of the online survey, you indicate your consent to participate in this study.

Annex 21: Mail Invitation

Dear <first name> <last name>,

GESIS wants to improve its data services and make them more user-friendly. You can support us by completing a survey on your experience with searching and using survey data.

Participating in this survey will take about 10 to 15 minutes. Your responses will be collected anonymously with an online survey system. You can access the survey here:
https://www.1ka.si/a/183652?group=13728520

This research is part of my PhD project on searching and reusing survey data that I am conducting at GESIS and at the Humboldt University Berlin. Results of this study will be made available to a larger scientific community. That way, your participation in this survey may have broader impact on the development of research data infrastructure beyond GESIS services.

Further information on the research project and the data handling and data processing is available at http://dbk.gesis.org/data_search/ConsentForm.pdf

Thank you very much for supporting this research. Feel free to contact me with any questions you might have about this study.

Best regards,

Tanja Friedrich, M.A.
GESIS Leibniz Institute for the Social Sciences
Unter Sachsenhausen 6-8
50667 Köln
tanja.friedrich@gesis.org

Please note: You are receiving this personal e-mail, because you have agreed to receive further information and news related to the data catalogue and other GESIS services when creating your GESIS data catalogue account. You can revoke your agreement by changing the preferences in your data catalogue account at any time:
https://dbk.gesis.org/dbksearch/login.asp?db=e
For further information regarding the handling of your personal data, please visit:
https://www.gesis.org/en/institute/data-protection/

Annex 22: Mail Reminder

Dear <first name> <last name>,

On <date>, I invited you to participate in our survey on data search and data use.

I would like to thank you very much, if you have participated in this survey.

If you have not participated yet, you can still do so until <date>. You will find the link to the survey in the e-mail attached below.

I will not send you any further reminders.

Best regards,

Tanja Friedrich, M.A.
GESIS Leibniz Institute for the Social Sciences
Unter Sachsenhausen 6-8
50667 Köln
tanja.friedrich@gesis.org

Dear <first name> <last name>,

GESIS wants to improve its data services and make them more user-friendly. You can support us by completing a survey on your experience with searching and using survey data.

Participating in this survey will take about 10 to 15 minutes. Your responses will be collected anonymously with an online survey system. You can access the survey here: https://www.1ka.si/a/183652?group=13728520

This research is part of my PhD project on searching and reusing survey data that I am conducting at GESIS and at the Humboldt University Berlin. Results of this study will be made available to a larger scientific community. That way, your participation in this survey may have broader impact on the development of research data infrastructure beyond GESIS services.

Further information on the research project and the data handling and data processing is available at http://dbk.gesis.org/data_search/ConsentForm.pdf

Thank you very much for supporting this research. Feel free to contact me with any questions you might have about this study.

Best regards,

Tanja Friedrich, M.A.

GESIS Leibniz Institute for the Social Sciences
Unter Sachsenhausen 6-8
50667 Köln
tanja.friedrich@gesis.org

Please note: You are receiving this personal e-mail, because you have agreed to receive further information and news related to the data catalogue and other GESIS services when creating your GESIS data catalogue account. You can revoke your agreement by changing the preferences in your data catalogue account at any time:
https://dbk.gesis.org/dbksearch/login.asp?db=e
For further information regarding the handling of your personal data, please visit:
https://www.gesis.org/en/institute/data-protection/

Annex 23: Pop-up Text (Web Survey)

Pop up invitation for DBK users

English text:

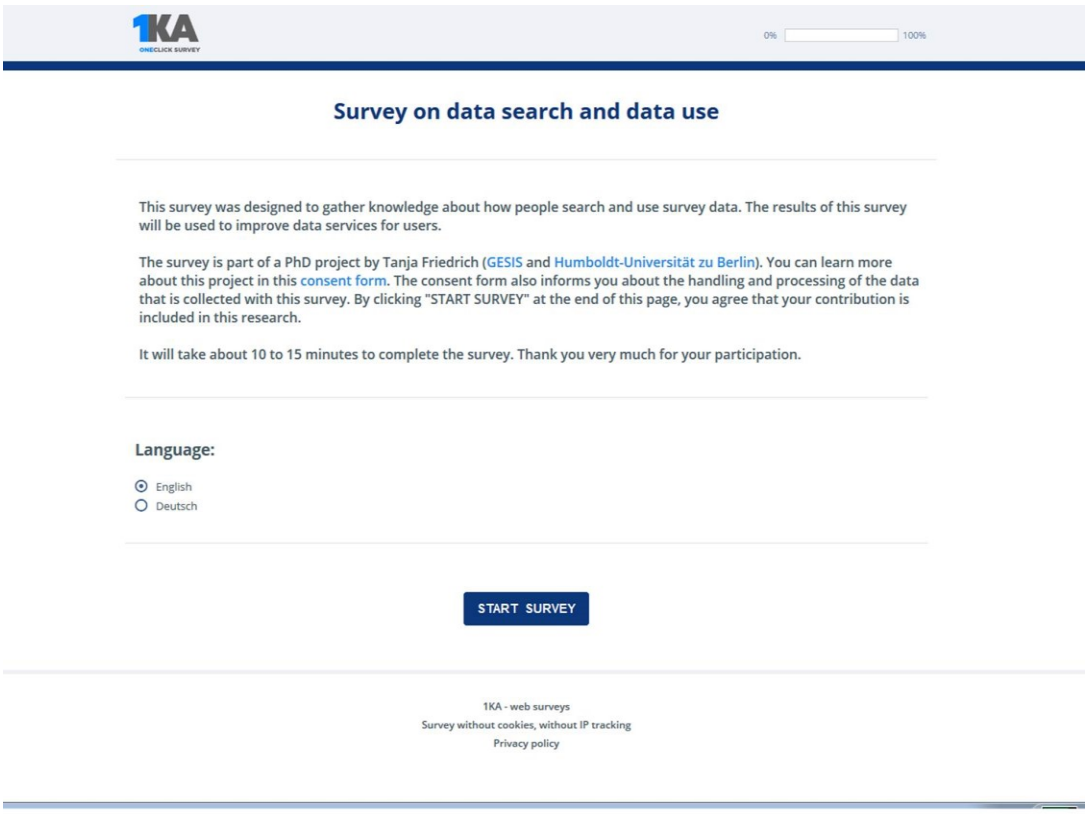
We are conducting a survey on data search and data use. If you would like to help us improve our data services, you can access the survey here: <Participate> (https://www.1ka.si/a/183652?group=13728521).

German text:

Wir führen gerade eine Umfrage zur Datensuche und Datennutzung durch, um unsere Services zu verbessern. Wären Sie bereit, uns dabei zu helfen? Bitte klicken Sie hier, um an der Umfrage teilzunehmen: <Teilnehmen> (https://www.1ka.si/a/183652?group=13728521&language=5).

Annex 24: Web Survey Start Page

English version:



German version: